T E S I D I D O T T O R A T O

Dipartimento di Scienze Economiche Aziendali e Statistiche

Modello di regressione esteso per l'analisi semiparametrica dei dati di sopravvivenza

Extended Hazard Regression for Survival Data

Rossella Alduino

Tutor: *Prof. Vito Muggeo*

Coordinatore Dottorato: *Prof. Marcello Chiodi*

Dottorato di Ricerca in "Statistica, Statistica Applicata e Finanza Quantitativa, XXV Ciclo
Settore Scientifico Disciplinare: SECS/S01 - Statistica
Anno conseguimento titolo: 2016

Università degli Studi di Palermo

DSEAS
Dipartimento di Scienze Economiche, Aziendali e Statistiche

A Gigi e Fabio

Ringraziamenti

Desidero ringraziare tutti coloro che mi hanno dato la possibilitá di realizzare questa tesi e di concludere, cosí, un importante percorso di studi.
Ringrazio innanzitutto il prof. Vito Muggeo per aver accettato di seguirmi con interesse e professionalitá in questo lavoro, per la sua immensa disponibilitá, per avermi dedicato buona parte del suo tempo e per avermi fornito preziosi aiuti, suggerimenti e consigli.
Ringrazio tutti i docenti del dipartimento, primo fra tutti il Prof. Marcello Chiodi, per la loro disponibilitá e per avermi dato la possibilitá di frequentare corsi e seminari che hanno sicuramente contribuito all'arricchimento delle le mie conoscenze e competenze.
Ringrazio Nicola e i miei genitori per esserci sempre, per sostenermi in ogni momento e per aver sempre fiducia in me e nelle mie decisioni.
Ringrazio, ancora Nadia, Giuseppe, Clara e tutti gli amici e colleghi che mi hanno sostenuta moralmente ed incoraggiata affinché potessi raggiungere questo traguardo cosí importante.
Infine, un ringraziamento particolare ai miei piccoli Gigi e Fabio che rendono speciale la mia vita.

Indice

Elenco delle figure

Elenco delle tabelle

Introduzione

Con il termine *Analisi della sopravvivenza* si fa riferimento ad una serie di metodi e modelli statistici finalizzati all'analisi di dati caratterizzati da tempi che intercorrono tra un istante 0, solitamente l'inizio dello studio, fino al verificarsi di un determinato evento che è detto *end-point* (Collett, 2003; Kleinbaum and Klein, 2005; Marubini and Valsecchi, 1995).
L'analisi della sopravvivenza puó essere applicata a diversi campi, come medicina, sanitá pubblica, scienze sociali, ingegneria. Ad esempio in medicina il tempo di accadimento puó essere il tempo fino alla morte del paziente o il tempo fino al verificarsi di un'infezione (Tableman and Kim, 2004). Nelle scienze sociali, l'interesse potrebbe essere l'analisi dei tempi di accadimento di un evento quale, per esempio, il cambiamento di lavoro, il matrimonio, la nascita di un bambino e cosí via (Lee and Wang, 2003). In tutti questi campi di applicazione il termine maggiormente utilizzato è analisi della sopravvivenza, anche se nelle diverse discipline, spesso vengono utilizzati differenti nomi, ma ció non implica una reale differenza nelle tecniche d'analisi. Per esempio, in sociologia a volte si parla di *event history analysis*, in economia di *duration* o *transition analysis* ed in campo ingegneristico, dove il principale obiettivo è quello di valutare i tempi di vita di macchinari o di componenti elettronici, si parla di *failure time analysis*

(Kalbfleisch and Prentice, 2002).

L'obiettivo di questa tesi è quello di provare a studiare ed analizzare i dati di sopravvivenza attraverso l'utilizzo di un modello semiparametrico generale. A tale scopo nei capitoli seguenti ci sará una descrizione dell'analisi della sopravvivenza e delle principali funzioni che vengono utilizzate; ci sará, altresí una panoramica dei metodi statistici utilizzati, siano essi non parametrici, parametrici o semiparametrici. Particolare attenzione sará posta a quest'ultimi per la loro caratteristica principale: quella di non fare alcuna assunzione distribuzionale sui tempi di sopravvivenza.
I modelli semiparametrici utilizzati per l'analisi dei dati di sopravvivenza che sono discussi in questa tesi sono elencati di seguito.

- I modelli a rischi proporzionali, dei quali l'approccio piú comunemente utilizzato è quello introdotto da Cox (1972);

- I modelli a tempi di evento accelerati o modelli AFT (*Accelerated Failure Time*): sono modelli generali per dati di sopravvivenza, in cui si assume che le variabili esplicative misurate su un soggetto agiscano moltiplicativamente sulla scala temporale e cosí hanno il ruolo di aumentare o diminuire la velocitá con cui un individuo procede lungo l'asse dei tempi, accelerando o rallentando il verificarsi dell'evento terminale (Collett, 2003, 195-200).

- I modelli a rischi accelerati o modelli AH (*Accelerated Hazard*): particolarmente utili per modellare l'effetto di un trattamento o di altre esplicative sulla distribuzione dei tempi di sopravvivenza, quando l'effetto del trattamento è graduale e c'è un ritardo prima che il trattamento sia completamente efficace (Chen *et al.*, 2014).

Il modello AH è un modello alternativo per l'analisi di dati di sopravvivenza. Tuttavia, la complessitá dei metodi di stima semiparametrica esistenti in letteratura, ostacola la sua applicazione.

Lo scopo di questa tesi è quello di descrivere un modello semiparametrico esteso (EH, extended hazard) che comprenda come casi particolari il modello PH di Cox, il modello AFT e quello AH e che possa essere stimato in modo relativamente semplice attraverso la stima iterativa di opportune regressioni log lineari di Poisson.
L'implementazione di un modello di Poisson in pratica richiede che il *follow-up* per ciascun soggetto sia suddiviso in piccoli intervalli di tempo. Conseguentemente le esplicative *time-varying* potranno essere inserite per ciascun intervallo, mentre quelle fisse nel tempo (ed es., genere) verranno ripetute per ciascun soggetto, in tutti gli intervalli. I dati organizzati in questa maniera, consentono di fare una chiara distinzione tra il *risk time* che è la lunghezza di ciascun intervallo e il *time-scale* che è il valore del tempo all'inizio di ciascun intervallo (Carstensen, 2005).

La tesi è articolata in sei capitoli. Il Capitolo 1 fornisce una panoramica dell'analisi della sopravvivenza e passa in rassegna i modelli parametrici e non parametrici di regressione. Il Capitolo 2 si concentra sull'analisi dei modelli semi-parametrici, di cui fa parte il modello di Cox a rischi proporzionali, il modello AFT e il modello AH. Il Capitolo 3 fornisce un confronto tra i diversi modelli attraverso la descrizione di una classe generale di modelli di regressione semiparametrica basati sulla regressione di Poisson e sulla sua possibile estensione all'analisi dei dati di sopravvivenza. Il Capi-

tolo 4 mostra la procedura di stima del modello esteso. Il Capitolo 5 fornisce i risultati degli studi di simulazione. Il Capitolo 6 mostra l'applicazione del modello esteso a dati reali.

Capitolo 1

Analisi della Sopravvivenza

Con il termine *Analisi della sopravvivenza* si fa riferimento ad una Tecnica statistica di analisi di dati, ottenuti da una coorte di unitá osservate longitudinalmente, che consente di stimare la probabilitá del verificarsi di un determinato evento in funzione del tempo (Collett, 2003, 1-3).
Nelle prime applicazioni l'evento in studio era quasi sempre la morte del paziente; da ció il nome *dati di sopravvivenza*. Successivamente questo termine venne utilizzato per tutti i tipi di eventi in studio (Lee and Wang, 2003). In ogni caso l'evento puó essere considerato come una transizione da uno stato ad un altro.
Negli studi clinici, spesso, il tempo d'origine corrisponde all'inserimento di un individuo in uno studio sperimentale, come una prova clinica per confrontare due o piú trattamenti, altrimenti puó coincidere con la diagnosi relativa ad una particolare condizione, l'inizio di un trattamento, o il verificarsi di qualche evento sfavorevole (Tableman and Kim, 2004, 2-3). Per esempio, bambini sottoposti ad un trapianto di rene possono essere seguiti per identificare eventuali predittori della mortalitá. In particolare, il rischio

di mortalitá è piú basso per quei bambini che ricevono il rene da un donatore vivente? È possibile che il rischio di mortalitá dipenda dal tempo impiegato per il trasporto del rene del donatore? Quanto incide sulla sopravvivenza del paziente trapiantato, la corrispondenza tra le caratteristiche del donatore e quelle del ricevente? In tutti gli studi l'interesse comune è la descrizione della relazione tra una o piú variabili di esposizione e il *tempo di sopravvivenza*. Durante il periodo di osservazione, peró, solo alcuni individui sperimentano l'evento finale, mentre altri soggetti alla fine del *follow up* non hanno ancora sperimentato l'evento, pertanto non si conosce il loro tempo di sopravvivenza. Da ció la presenza di dati censurati (Selvin, 2008, 73).
Esistono tre tipi di censure: censura a destra, censura a sinistra e intervallo censurato. La censura a destra è molto comune nei dati di sopravvivenza, mentre la censura a sinistra è piuttosto rara. Il termine censura sará usato in questa tesi per indicare esclusivamente le censure a destra.

La censura a destra avviene dopo che l'individuo è entrato nello studio, cioé alla destra dell'ultimo tempo di sopravvivenza noto. Il tempo di sopravvivenza censurato a destra è quindi minore dell'effettivo, ma sconosciuto, tempo di sopravvivenza. Ci sono diverse ragioni per le quali si puó osservare una censura a destra: per esempio, al termine dello studio l'evento considerato non si è ancora verificato, oppure il soggetto si ritira dallo studio o esce dallo studio per ragioni diverse. Se per esempio l'*end point* è la morte del paziente, un tempo di sopravvivenza puó essere ritenuto censurato a destra quando la morte è causata da motivi non legati al trattamento o al fenomeno in studio. La sola informazione utile sull'esperienza relativa alla sopravvivenza di questo paziente è l'ultima data in cui si sa essere an-

cora vivo.

In sintesi, l'analisi della sopravvivenza si basa sullo studio di dati del tipo (T, C), dove si osserva soltanto $min(T, C)$ e $\delta = 1\{T < C\}$: T è il tempo di sopravvivenza, C il tempo di censura e δ è la variabile indicatrice di censura (Kleinbaum and Klein, 2005, 8).

Nell'analisi statistica della sopravvivenza bisogna considerare due aspetti importanti: il primo è connesso con il fatto che la gran parte delle funzioni di densitá di probabilitá della variabile casuale T è fortemente asimmetrica positiva e spesso non si ha un'informazione esaustiva sulla distribuzione della variabile casuale T . Il secondo aspetto è la presenza di dati censurati che rende gli usuali metodi statistici inappropriati per l'analisi (Marubini and Valsecchi, 1995).

1.1 Distribuzione del tempo di sopravvivenza

Nell'analizzare i dati di sopravvivenza le funzioni di interesse centrale sono: la funzione di sopravvivenza, la funzione di densitá di probabilitá e la funzione di rischio riferita talvolta con il termine anglosassone *hazard function*.
L'effettivo tempo t di sopravvivenza di un individuo, puó essere considerato come il valore di una variabile aleatoria T, che assume solo valori non negativi. Si supponga che la variabile casuale T abbia una distribuzione di probabilitá $F(t)$ con *funzione di densitá di probabilitá* $f(t)$. La distribuzione di probabilitá di T è data da:

$$F(t) = P(T < t) = \int_0^t f(u)du, \tag{1.1}$$

e rappresenta la probabilitá che il tempo di sopravvivenza sia inferiore ad un dato valore t.
La *funzione di sopravvivenza*, $S(t)$, è definita come la probabilitá che il tempo di sopravvivenza T sia maggiore o uguale al valore t:

$$S(t) = P(T \geq t) = 1 - F(t) = 1 - P(T < t). \tag{1.2}$$

La funzione di sopravvivenza puó quindi essere usata per rappresentare la probabilitá che un individuo sopravviva oltre il tempo t.
La funzione di rischio è la probabilitá che un individuo muoia al tempo t, condizionata al fatto che sia sopravvissuto fino a quell'istante:

$$\lambda(t) = \lim_{\Delta t \to 0} \left\{ \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \right\}. \tag{1.3}$$

La funzione di rischio quindi rappresenta la velocitá istantanea dell'evento in studio per un individuo che non ha sperimentato l'evento fino al tempo t.
C'è una relazione tra la funzione di sopravvivenza $S(t)$ e la funzione di rischio $\lambda(t)$, espressa dalla formula seguente:

$$\lambda(t) = \frac{f(t)}{S(t)} = -\frac{d}{dt} \log S(t) \tag{1.4}$$

$$S(t) = \exp\left\{ -\int_0^t \lambda(u) du \right\} = \exp\{-\Lambda(t)\}, t \geq 0. \tag{1.5}$$

dove $\Lambda(t) = \int_0^t \lambda(u) du$ è chiamata *funzione di rischio cumulato*, la quale puó essere ottenuta dalla funzione di sopravvivenza poiché $\Lambda(t) = -\log S(t)$. La funzione di densitá di probabilitá T puó essere scritta quindi nel seguente modo:

$$f(t) = \lambda(t) \exp\left\{ -\int_0^t \lambda(u) du \right\}, t \geq 0 \tag{1.6}$$

Queste tre funzioni forniscono un'equivalente specificazione della distribuzione del tempo di sopravvivenza T. Se si conosce una delle tre funzioni, le

altre due si possono facilmente determinare. (Kleinbaum and Klein, 2005, 11). È sufficiente, quindi, sceglierne una come base per l'analisi statistica a seconda degli obiettivi finali dello studio. La funzione di sopravvivenza è maggiormente utile per confrontare la sopravvivenza di due o piú gruppi di soggetti; la funzione di rischio fornisce una descrizione del rischio di sperimentare l'evento in ogni istante temporale osservato.

1.2 Modelli di sopravvivenza

I metodi dell'analisi della sopravvivenza possono essere classificati in metodi non parametrici, metodi semi-parametrici e metodi parametrici, sulla base degli assunti che vengono fatti sulla distribuzione di T. In questo paragrafo verranno discussi i metodi di stima della funzione di sopravvivenza parametrici e non parametrici; nel capitolo successivo si descriveranno i metodi semi-parametrici.

Negli scenari di vita reale, spesso, non si conosce l'esatta distribuzione dei dati, pertanto ogni qualvolta non si è in grado di fare ipotesi distribuzionali sui dati di sopravvivenza in esame, è opportuno trattare tali dati in forma non parametrica.

I metodi non parametrici, sono molto semplici da comprendere e da applicare; sono meno indicati dei metodi parametrici quando i tempi di sopravvivenza seguono una distribuzione teorica e maggiormente adatti quando non si conosce tale distribuzione (Tableman and Kim, 2004, 25). Ovviamente l'applicazione dei metodi non parametrici è principalmente indicata quando l'obiettivo è quello di fare un'analisi esplorativa dei dati.

1.2.1 Modelli di sopravvivenza non parametrici

Il metodo non parametrico piú diffuso per la stima della probabilitá di sopravvivenza è il metodo del prodotto limite, noto anche come stimatore di Kaplan-Meier (Kaplan and Meier, 1958). Consiste nello stimare la probabilitá condizionata di sopravvivenza in corrispondenza di ciascuno dei tempi in cui si verifica almeno un evento terminale.

Lo stimatore di Kaplan-Meier della funzione di sopravvivenza è cosí definito:

$$\hat{S}(t) = \prod_{y_{(i)} \leq t} \hat{p}_i = \prod_{y_{(i)} \leq t} \frac{n_i - d_i}{n_i} \tag{1.7}$$

dove $y_{(k)} \leq t < y_{(k+1)}$
essendo:

- n_i il numero dei soggetti a rischio prima di $y_{(i)}$;
- d_i numero di soggetti che sperimentano l'evento al tempo $y_{(i)}$;
- $p_i = P(T > y_{(i)} | T > y_{(i-1)})$

A livello teorico, se si considera il tempo $t \in (0, +\infty)$, la funzione di sopravvivenza puó essere rappresentata come una curva a gradini, che parte dal valore $S(0) = 1$ e decresce nel tempo $S(\infty) = 0$. La stima di Kaplan-Meier della probabilitá di sopravvivenza viene rappresentata con una curva a gradini, che parte dal valore 1 e decresce nel tempo. L'altezza dei gradini dipende dal numero di eventi e dal numero di soggetti a rischio. Si veda la Figura 1.1 per un esempio.

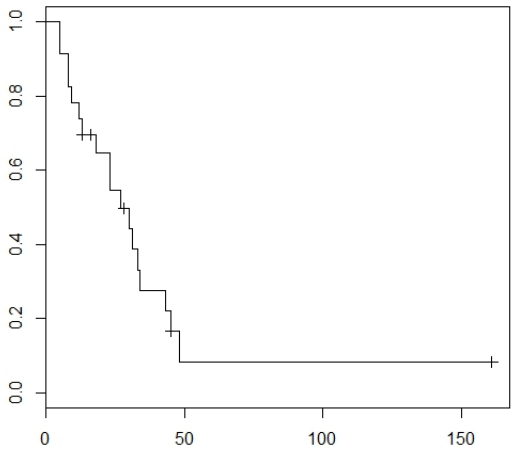

Figura 1.1: Esempio di curva di Kaplan-Meier

Un altro metodo non parametrico per la stima della funzione di sopravvivenza è il metodo di calcolo delle tavole di sopravvivenza, conosciuto anche come metodo attuariale (Collett, 2003, 17). Tale metodo consiste nella suddivisione dell'asse temporale in intervalli di tempo, ad esempio in numero $I + 1$, di uguale ampiezza, tranne l'ultimo che ha ampiezza infinita. Su un campione casuale di n soggetti, vengono rilevati i tempi di sopravvivenza e si contano quanti individui presentano l'evento d'interesse, o l'osservazione censurata, in ciascun intervallo. Gli elementi necessari alla costruzione della tavola di sopravvivenza sono quindi (per $i = 1, ..., I + 1$):

- l'ampiezza dell'intervallo i-esimo;
- il numero di individui, n_i, che entrano nell'intervallo;

- il numero di eventi, d_i, verificatisi nell'intervallo.

È possibile definire una stima della probabilitá condizionata che un soggetto sperimenti l'evento nell'intervallo i-esimo, dato che non l'ha sperimentato nell'intervallo immediatamente precedente: $\frac{d_i}{n_i}$. La stima della probabilitá di sopravvivenza nell'intervallo i-esimo è data, ovviamente, da $1 - \frac{d_i}{n_i}$. Il prodotto delle stime delle probabilitá condizionate di sopravvivere in ciascuno degli intervalli fino allo i-esimo, definisce la probabilitá cumulativa di sopravvivenza.

A differenza della procedura adottata nel metodo attuariale, il metodo di Kaplan-Meier non implica la suddivisione dell'asse temporale in intervalli di ampiezza prefissata e quindi neppure il conseguente raggruppamento di soggetti.

1.2.2 Modelli di sopravvivenza parametrici

Un modello di sopravvivenza parametrico è un modello per il quale si assume che il tempo di sopravvivenza segua una distribuzione nota (Kleinbaum and Klein, 2005, 258-265).
I metodi parametrici per l'analisi dei dati di sopravvivenza consentono, quindi, la stima della funzione di sopravvivenza e della funzione di rischio, adattando ai dati una funzione di cui si assume una certa forma e che dipende da uno o piú parametri. Naturalmente, è necessario tenere presente che la loro applicabilitá è limitata dal fatto che la funzione di rischio assume una ben definita forma matematica, mentre in molte situazioni pratiche non vi sono sufficienti giustificazioni per l'adozione di un particolare modello.
Esempi di distribuzioni che sono comunemente utilizzate per dati di sopravvivenza sono: la distribuzione di Weibull, l'esponenziale (che è un caso

particolare della distribuzione di Weibull), la log-logistica, la log-normale e la gamma generalizzata. A causa della sua popolaritá e della semplicitá di calcolo oltre che delle importanti proprietá che possiede, la distribuzione esponenziale è uno dei modelli parametrici maggiormente utilizzati nell'analisi dei dati di sopravvivenza (Hosmer and Lemeshow, 1999, 275).

Si supponga di disporre di un campione di tempi di sopravvivenza t_i, $i = 1, ..., n$. È spesso d'interesse valutare se, e in che modo, la distribuzione di sopravvivenza è influenzata da variabili concomitanti $x = (x_{i1}, ..., x_{ip})$, dove x_{ir} è il valore assunto dalla r-esima variabile concomitante per l'i-esima unitá statistica. La modellazione piú semplice si ha specificando un modello parametrico per la distribuzione marginale dei tempi t_i.
Alcuni tra i modelli parametrici piú usati nell'analisi della sopravvivenza sono riportati in Tabella 1.1, insieme alla relativa funzione di rischio.
I modelli parametrici di regressione sono in genere distinti a seconda del comportamento della funzione di rischio.

Tabella 1.1: Funzioni di Sopravvivenza e di Rischio per tipo di distribuzione

Distribuzione	$S(t)$	$\lambda(t)$
Esponenziale	$exp(-\lambda t)$	λ
Weibull	$exp(-(\lambda t)^{\alpha})$	$\lambda\alpha(\lambda t)^{\alpha-1}$
Log-logistica	$\frac{1}{1+\lambda t^{\alpha}}$	$\frac{\lambda\alpha t^{\alpha-1}}{1+\lambda t^{\alpha}}$
Log-normale	$1 - \Phi(z)$	$\frac{(2\pi)^{-\frac{1}{2}}(\sigma t)^{-1}\exp[-\frac{1}{2}(z^2)]}{1-\Phi(z)}$

con $\Phi(z) = \int_{-\infty}^{Z}(2\pi)^{-\frac{1}{2}}\exp\frac{-u^2}{2}du$ e $z = \frac{\ln t-\mu}{\sigma}$

L'esponenziale è una distribuzione ad un parametro con funzione *hazard* costante e pari a λ. Il modello esponenziale presuppone, quindi, che il rischio istantaneo del verificarsi dell'evento non vari al trascorrere del tempo. Questo implica, per esempio in uno studio clinico, che il rischio di morte nel primo anno di osservazione del paziente in esame, è pari al rischio di morte in ciascuno degli anni successivi di osservazione. Un esempio dove questo non accade è dato dall'osservazione del periodo immediatamente successivo all'operazione chirurgica per i pazienti che sono stati sottoposti, per esempio ad un trapianto, periodo in cui si nota un eccesso di mortalitá dovuto al forte rischio di infezione o di rigetto. Il rischio di morte dopo questo periodo risulta essere piú piccolo. In generale, per distinguere la funzione di rischio costante nel tempo da quella che invece varia al variare del tempo, si usa la notazione λ in luogo di $\lambda(t)$.
Le distribuzioni Weibull e log-logistica hanno due parametri λ e α. La distribuzione di Weibull si riduce all'esponenziale quando $\alpha = 1$. Nella distribuzione di Weibull la funzione *hazard* è:

- per $\alpha < 1$ monotona decrescente;
- per $\alpha > 1$ monotona crescente;
- per $\alpha = 1$ costante e pari a λ.

Ovvero, il modello di Weibull presuppone che la funzione rischio possa essere costante nel tempo, monotona crescente o monotona decrescente, a seconda del valore assunto dal parametro di forma α della distribuzione (Hosmer and Lemeshow, 1999, 289).

Il modello log-normale, invece, è l'unico ad avere una funzione di rischio non monotona, ma che cresce dal valore iniziale nullo $t = 0$ fino a

raggiungere un punto di massimo, per poi decrescere verso zero per $t \to \infty$.

Solitamente nei modelli di sopravvivenza parametrici il parametro λ viene riparametrizzato in termini di predittori e di parametri della regressione e il parametro α (di solito definito parametro di forma) viene tenuto costante.

Osservando la Tabella 1.1 possiamo notare che la funzione di densitá di probabilitá per queste tre distribuzioni si ricava facilmente moltiplicando $\lambda(t)$ e $S(t)$ (Kleinbaum and Klein, 2005, 263).
Per esempio la funzione di densitá di probabilitá della Weibull è:

$$f(t) = \lambda\alpha t^{\alpha-1} exp(-\lambda t^{\alpha}) \tag{1.8}$$

poiché $\lambda(t) = \lambda\alpha t^{\alpha-1}$ e $S(t) = exp(-\lambda t^{\alpha})$.

Incorporando esplicitamente il vettore x delle esplicative nella formulazione parametrica della funzione rischio, è possibile studiare il loro effetto sulla sopravvivenza. Ad esempio, la funzione di rischio studiata da (Glasser, 1967), ottenuta assumendo che la variabile casuale T, dato il vettore x, sia distribuita in accordo ad un modello esponenziale, prevede:

$$\lambda(t; x) = \lambda \exp(x^T \beta), \tag{1.9}$$

dove β è un vettore di $p \times 1$ ignoti parametri di regressione, x è un vettore di esplicative e $\lambda > 0$.
I coefficienti di regressione β costituiscono una misura quantitativa dell'effetto esercitato da ciascuno dei p fattori prognostici sul rischio di morte. La stima di β, cosí come i test sui coefficienti di regressione, ricorrono alla

teoria della verosimiglianza.
Il modello di Weibull assume invece:

$$\lambda(t; x) = \lambda\alpha(\lambda t)^{\alpha-1} \exp(x^T\beta). \tag{1.10}$$

Se si considera la formulazione log-lineare la (1.9) diventa:

$$\log(\lambda(t; x)) = \alpha_0 + (x^T\beta), \tag{1.11}$$

cioé il logaritmo del rischio è funzione lineare di p esplicative. La costante α_0 rappresenta una log-baseline del rischio, ossia il rischio che si ha quando tutte le esplicative assumono valori uguali a zero.

Ne segue che il rapporto tra i rischi (*hazard ratio*) per due individui con vettori di esplicative, rispettivamente, x_1 e x_2 non dipende dal tempo poiché si ha:

$$HR = \frac{\lambda(t; x_1)}{\lambda(t; x_2)} = \frac{\exp(x_1^T\beta)}{\exp(x_2^T\beta)}. \tag{1.12}$$

Pertanto, l'effetto di ciascuna esplicativa sull'HR è moltiplicativo.

Per mezzo delle relazioni viste è possibile definire una piú ampia classe di modelli per i quali la funzione di rischio è scomponibile in due fattori, di cui uno dipendente solo dal tempo e l'altro solo dalle esplicative, ossia:

$$\lambda(t; x) = \lambda_0(t)h(x). \tag{1.13}$$

Quest'ultima equazione caratterizza la classe di modelli che presuppongono la proporzionalitá delle funzioni di rischio (*proportional hazard models*).

Si ottengono diversi modelli in base alle differenti forme parametriche assunte da $\lambda_0(t)$.
La scelta di $h(x)$ dipende dal tipo di dati sotto studio e non deve essere necessariamente di forma esponenziale.

Un'altra classe di modelli utili per l'analisi della sopravvivenza è costituita dai modelli con tempi di evento accelerati (*accelerated failure time models*) (Cox and Oakes, 1984, 64). In questa classe di modelli si assume che il vettore x agisca moltiplicativamente sul tempo di sopravvivenza: il ruolo di x consiste nell'aumentare o diminuire la velocitá con cui un individuo procede lungo l'asse dei tempi, accelerando o rallentando il verificarsi dell'evento terminale. Un modello AFT puó essere scritto come:

$$\log T_i = x_i^T \beta + \alpha\epsilon \tag{1.14}$$

in cui α ($\alpha > 0$) è un parametro di scala e ϵ è una variabile casuale che si assume abbia una particolare distribuzione.
In un approccio parametrico, per ogni distribuzione di ϵ, si ha una corrispondente distribuzione di T (Tabella1.2).

Tabella 1.2: Alcune densitá per il modello AFT

Distribuzione di ϵ	Distribuzione di T
Extreme value (1 parametro)	Esponenziale
Extreme value (2 parametri)	Weibull
Logistica	Log-logistica
Normale	Log-normale
Log-Gamma	Gamma

I modelli AFT sono stimati mediante il metodo della massima verosimiglianza (Collett, 2003, 216). La verosimiglianza di n tempi di sopravvivenza osservati $t_1, t_2, ..., t_n$ è data da:

$$L(\alpha, \beta) = \prod_{i=1}^{n} f_i(t_i)^{\delta_i} S_i(t_i)^{1-\delta_i} \tag{1.15}$$

dove $f_i(t_i)$ e $S_i(t_i)$ sono le funzioni di densitá e di sopravvivenza per l'i-esimo individuo al tempo t_i e δ_i è la variabile indicatrice dell'evento. Le stime di massima verosimiglianza dei parametri possono essere ottenute massimizzando la funzione di log-verosimiglianza attraverso metodi numerici come quello di Newton-Raphson.

Le classi di modelli parametrici sopra brevemente descritte, forniscono una varietá di metodi per l'analisi della sopravvivenza che non è esaurita da quelli presentati in Tabella 1.1. È necessario tenere presente che la loro applicazione dipende anche dagli obiettivi dello studio e dalle proprietá che i diversi modelli parametrici hanno. Cox and Oakes (1984, 65), per esempio, hanno dimostrato che i soli modelli a tempi di evento accelerati che hanno anche la proprietá degli hazard proporzionali sono i modelli esponenziale e Weibull (Hosmer and Lemeshow, 1999, 275)

Riassumendo possiamo asserire che i modelli parametrici consentono di perseguire due obiettivi simultaneamente. Il modello deve descrivere la distribuzione sottostante il tempo di sopravvivenza (componente dell'errore), ma deve anche caratterizzare i cambiamenti della distribuzione in funzione delle esplicative (componente sistematica) (Hosmer and Lemeshow, 1999, 271-273). A livello applicativo, in alcuni casi è importante usare un modello che persegua entrambi gli obiettivi, ma in altri casi è sufficiente

un modello che persegua soltanto il secondo degli obiettivi descritti. Supponiamo di voler sapere se una combinazione di terapie farmacologiche migliora la sopravvivenza di pazienti affetti da HIV rispetto ad una terapia farmacologica singola. In questo caso, una completa descrizione del tempo di sopravvivenza è di secondaria importanza rispetto alla descrizione di come la nuova terapia modifica la sopravvivenza di un paziente rispetto alla vecchia terapia. In questo esempio, è necessario stimare i parametri che possono essere usati per confrontare la sopravvivenza dei due gruppi di trattamento, e questo confronto puó essere fatto inserendo anche altre variabili, per esempio l'etá o il sesso del paziente. I modelli parametrici possono essere usati per perseguire tale obiettivo. Tuttavia, le assunzioni richieste per le loro componenti dell'errore possono risultare inutilmente rigorose, dato che l'inferenza riguarderá soltanto i parametri della parte sistematica del modello (Hosmer and Lemeshow, 1999, 89).
I modelli usati per descrivere i tempi di sopravvivenza in senso comparativo vengono chiamati modelli di regressione semi-parametrica e sono quelli di cui ci occuperemo prevalentemente in questa tesi.

Capitolo 2

Modelli semi-parametrici di regressione

Nel precedente capitolo abbiamo evidenziato che è possibile descrivere la distribuzione del tempo di sopravvivenza in uno di due equivalenti modi. Possiamo specificare la funzione di densitá di una distribuzione parametrica o possiamo specificare la funzione di rischio. In ogni caso, la specificazione di un modello deve dare la possibilitá di rispondere a specifiche domande quali, per esempio, in che modo la sopravvivenza è correlata al tipo di trattamento in studio o ad altre caratteristiche dei pazienti (Hosmer and Lemeshow, 1999, 90-93).
Il modo naturale di cominciare è quello di considerare un modello di regressione sulla funzione di rischio, specificandola come funzione del tempo e delle esplicative.

$$\lambda(t, x, \beta) = \lambda_0(t) r(x, \beta) \tag{2.1}$$

dove $\lambda_0(t)$ rappresenta la funzione di rischio per i soggetti per i quali tutte le variabili esplicative hanno valore nullo. Quindi $\lambda_0(t)$ è la funzione

di rischio quando $r(x,\beta) = 1$ ed è definita funzione *baseline hazard*. La funzione di rischio, cosí come espressa nella formula, è il prodotto di due funzioni. La funzione $\lambda_0(t)$, indica come il rischio cambia in funzione del tempo di sopravvivenza e non coinvolge il vettore delle esplicative. L'altra funzione, $r(x,\beta)$, indica come la funzione di rischio cambia in funzione delle esplicative, ma non dipende dal tempo.
Cox (1972) fú il primo a proporre questo tipo di modello suggerendo di utilizzare:
$r(x,\beta) = exp(x\beta)$. Con questa parametrizzazione la funzione di rischio diventa:

$$\lambda(t,x,\beta) = \lambda_0(t)e^{x^T\beta} \tag{2.2}$$

Ovvero nella formulazione log-lineare:

$$\log(\lambda(t,x,\beta)) = \alpha_0(t) + x^T\beta \tag{2.3}$$

dove la baseline di rischio $\alpha_0(t)$ puó essere una qualsiasi funzione del tempo, mentre le esplicative sono in relazione lineare. Il modello per tale motivo si dice semi-parametrico poiché non descrive esplicitamente la funzione *hazard baseline* $\lambda_0(t)$.

2.1 Modello a rischi proporzionali

L'approccio piú comunemente utilizzato per modellare la distribuzione del tempo di sopravvivenza rispetto ad un insieme di variabili esplicative è il modello a rischi proporzionali introdotto da Cox (1972).
I dati basati su un campione di dimensione n, sono del tipo: (t_i, δ_i, x_i), $i = 1,\ldots,n$ dove t_i e il tempo per l'individuo i , δ_i è la variabile indicatrice dell'evento ($\delta_i = 1$ se per l'individuo i l'evento in studio si è verificato

e $\delta_i = 0$ se per il soggetto i si ha una censura) e x_i è il vettore di esplicative o di fattori di rischio per l'individuo i supposte avere un effetto sulla distribuzione del tempo di sopravvivenza T. La relazione tra il tempo di rischio e le esplicative x (dove x è un vettore di dimensione $1 \times p$) puó essere descritta attraverso il modello di Cox in cui la funzione *hazard* al tempo t per un individuo è:

$$\lambda_i(t, x) = \lambda_0(t)\, e^{x_i^T \beta}. \tag{2.4}$$

La formula del modello di Cox, esprime la funzione hazard al tempo t come prodotto tra, $\lambda_0(t)$, *funzione hazard baseline*, che è funzione di t, ma non coinvolge il vettore delle esplicative e $\exp(x_i^T \beta)$ dove β è un vettore di parametri ignoti di dimensione $p \times 1$ (Collett, 2003, 111-120).
Il risultato è una sequenza di curve, proporzionali tra loro in ragione del fattore moltiplicativo $\exp(x_i^T \beta)$: ognuna di esse riporta i valori della funzione di rischio istantaneo per una diversa combinazione dei valori assegnati alle variabili esplicative. Il vantaggio del modello di Cox è che non viene fatta alcuna assunzione su $\lambda_0(t)$, infatti si parla di approccio semi-parametrico. Il punto di forza del modello di Cox è proprio la sua flessibilitá nel non definire la funzione $\lambda_0(t)$.
Infine, il modello di Cox appena descritto viene definito a rischi proporzionali (PH) poiché il rapporto tra le funzioni *hazard* calcolate per due specifici individui con valori dell'esplicativa pari rispettivamente a x e x^* è:

$$HR = \frac{\lambda(t|x)}{\lambda(t|x^*)} = \exp[(x - x^*)\beta], \tag{2.5}$$

un'espressione che non dipende da t.
L'assunzione PH implica, pertanto, che la quantitá HR sia costante nel tempo e in pratica significa che l'hazard per un dato individuo è proporzionale all'hazard di ogni altro individuo, dove la proporzionalitá costante s'inten-

de indipendente dal tempo.
La Figura 2.1 mostra a livello esemplificativo il logaritmo dell'hazard calcolato per due specifici gruppi di soggetti (gruppo 0 e gruppo 1), in un modello di Cox a rischi proporzionali.

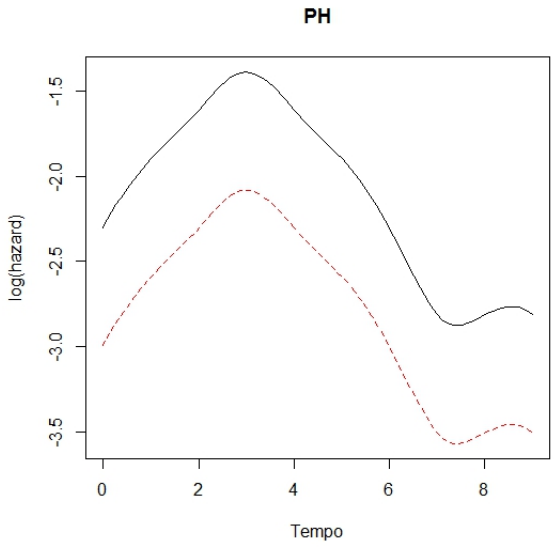

Figura 2.1: Log-hazard in un modello PH

Poiché il modello di Cox a rischi proporzionali prevede che l'effetto di una o piú esplicative non dipenda dal tempo, è molto importante verificare che tali esplicative soddisfino effettivamente l'assunzione di proporzionalitá dei rischi. Tuttavia tale assunzione è fortemente limitativa e non sempre verificabile su dati reali e la sua assunzione in molti casi è forzata.
L'approccio modellistico comporta la scelta di alcune variabili esplicative, ovvero è opportuno in genere seguire delle procedure che permettano di individuare fra le variabili a disposizione e le loro interazioni, quali siano

significative, ovvero quali possano essere inserite all'interno del modello, sulla base di un livello di significativitá. Una volta scelto il modello, questo deve essere stimato, ovvero è necessario stimare i coefficienti β e la funzione di rischio baseline $\lambda_0(t)$. Supponendo di disporre di n individui, m distinti tempi in cui si verifica l'evento in studio e $n - m$ tempi censurati. La funzione di verosimiglianza, vista come funzione dei parametri non noti del modello, proposta da Cox (1972) è pari a:

$$L(\beta) = \prod_{i=1}^{m} \frac{\exp(x_{(i)}^T \beta)}{\sum_{l \in R_{(t_{(i)})}} \exp(x_{(l)}^T \beta)} \tag{2.6}$$

dove $R_{(t_{(i)})}$ è l'insieme di individui a rischio al tempo $t_{(i)}$. Tale formula è definita *funzione di verosimiglianza parziale*. Il termine parziale è usato perché il modello di Cox considera la probabilitá soltanto per quei soggetti che sperimentano l'evento in studio (infatti si tratta di una verosimiglianza stimata a posteriori) e perché non considera la probabilitá per quei soggetti che sono stati censurati. Inoltre, vengono dapprima stimati i β e successivamente la baseline $\lambda_0(t)$.
Le stime di massima verosimiglianza dei parametri β possono essere ottenute massimizzando la funzione di log-verosimiglianza attraverso metodi numerici come quello di Newton-Raphson.

2.2 Modello a tempi di evento accelerati

Sebbene il modello a rischi proporzionali (PH) trovi grande applicabilitá nell'analisi dei dati di sopravvivenza, i modelli a tempi di evento accelerati o modelli AFT (*Accelerated Failure Time*) sono un'alternativa in circostanze dove non si puó considerare valida l'assunzione di proporzionalitá dei rischi (Orbe *et al.*, 2002), (Tableman and Kim, 2004, 101-105).

Il modello AFT è un modello generale per dati di sopravvivenza, in cui si assume che le variabili esplicative misurate su un soggetto agiscano moltiplicativamente sulla scala temporale e cosí hanno il ruolo di aumentare o diminuire la velocitá con cui un individuo procede lungo l'asse dei tempi, accelerando o rallentando il verificarsi dell'evento terminale (Collett, 2003, 195-200) (Kalbfleisch and Prentice, 2002, 240).
Il modello puó essere, quindi, interpretato in termini di velocitá di progressione di una malattia.

Per illustrare l'idea sottostante l'assunzione dei modelli AFT, si consideri un esempio. Supponiamo che dei pazienti siano casualmente assegnati ad uno di due trattamenti, quello standard, 0 e quello nuovo, 1 (Cox and Oakes, 1984, 64). Considerando un modello AFT, il tempo di sopravvivenza per un individuo sottoposto al nuovo trattamento puó essere pensato come multiplo del tempo di sopravvivenza di un individuo sottoposto al trattamento standard. Se T_0 rappresenta il tempo di sopravvivenza per l'individuo assegnato al trattamento standard e T_1 rappresenta il tempo di sopravvivenza per l'individuo assegnato al nuovo trattamento, allora l'assunzione del modello AFT puó essere espressa come segue:

$$T_1 = \phi T_0. \tag{2.7}$$

L'assunzione sottostante il modello AFT puó anche essere espressa in termini di funzione di sopravvivenza piuttosto che in termini di tempo di sopravvivenza. Considerando ció, la probabilitá che un individuo assegnato al nuovo trattamento sopravviva oltre il tempo t è uguale alla probabilitá che un individuo assegnato al trattamento standard sopravviva oltre il tempo t/ϕ, dove ϕ è una costante positiva.

Indichiamo con $S_0(t)$ e con $S_1(t)$ le funzioni di sopravvivenza per gli individui appartenenti ai due gruppi di trattamento. Allora il modello AFT specifica che:

$$S_1(t) = S_0(t/\phi), \tag{2.8}$$

per i valori del tempo di sopravvivenza t.

Un'interpretazione di questo modello è che la durata della vita di un individuo sottoposto al nuovo trattamento è ϕ volte la durata della vita di un individuo sottoposto al trattamento standard. Il parametro ϕ, quindi, riflette l'impatto del nuovo trattamento sulla scala del tempo (Zeng and Lin, 2007). Quando in uno studio, per esempio, l'*end-point* è la morte del paziente, valori di ϕ inferiori all'unitá corrispondono ad un'accelerazione del tempo di morte di un'individuo assegnato al nuovo trattamento, rispetto ad un individuo sottoposto al trattamento standard. Di contro, valori di ϕ superiori all'unitá corrispondono ad una decelerazione del tempo di morte di un'individuo assegnato al nuovo trattamento, rispetto ad un individuo sottoposto al trattamento standard.

La quantitá ϕ è definita *fattore di accelerazione* e rappresenta la misura dell'associazione ottenuta in un modello AFT ossia il *time ratio* (Hosmer and Lemeshow, 1999, 274). Questo fattore consente di valutare l'effetto delle esplicative sul tempo di sopravvivenza cosí come l'*hazard ratio* consente di valutare l'effetto delle esplicative sul rischio.

Il *time ratio* (TR) confronta due livelli dell'esplicativa X, $(x = 1 \; vs \; x = 0)$, tenendo costanti tutte le altre esplicative. Esso viene interpretato come il rapporto stimato dei tempi di sopravvivenza attesi per i due gruppi.

$TR > 1$ implica che l'esplicativa prolunga il tempo fino al verificarsi dell'evento in studio, ovvero l'esplicativa *accelera* il tempo di sopravvivenza; $TR < 1$ indica che è piú probabile che l'evento si verifichi prima, ovvero

l'esplicativa *decelera* il tempo di sopravvivenza.

Il fattore di accelerazione puó anche essere interpretato in termini di tempi di sopravvivenza mediani dei pazienti assegnati ai due trattamenti: $t_1(50)$ e $t_0(50)$. Questi valori sono tali che $S_1\{t_1(50)\} = S_0\{t_0(50)\} = 0.5$. Quindi, sotto il modello AFT, $S_1\{t_1(50)\} = S_0\{t_0(50)/\phi\}$, e cosí segue che $t_1(50) = \phi t_0(50)$. In altri termini, sotto un modello AFT, il tempo mediano di sopravvivenza di un paziente sottoposto al nuovo trattamento è ϕ volte quello di un paziente assegnato al trattamento standard.

$$T_1 = \phi T_0 \implies Me_1 = \phi Me_0 \tag{2.9}$$

dove Me_0 rappresenta il tempo di sopravvivenza mediano per un individuo appartenente al gruppo 0 e Me_1 rappresenta il tempo di sopravvivenza mediano per un individuo appartenente al gruppo 1.

Lo stesso ragionamento puó essere fatto per qualunque percentile della distribuzione del tempo di sopravvivenza. Questo significa che il q_{-esimo} percentile della distribuzione del tempo di sopravvivenza per un paziente assegnato al nuovo trattamento, $t_1(q)$, è tale che $t_1(q) = \phi t_0(q)$, dove $t_0(q)$ è il q_{-esimo} percentile per il trattamento standard.

La Figura 2.2 mostra a livello esemplificativo il logaritmo dell'hazard calcolato per due specifici gruppi di soggetti (gruppo 0 e gruppo 1), in un modello AFT.

In genere si assume che $\phi = e^{\alpha x}$.
Facendo tale assunzione e calcolando il log della (2.7) si ottiene:

$$\log T_1 = \log\{e^{x_i^T \alpha} T_0\} = \log T_0 + x_i^T \alpha \tag{2.10}$$

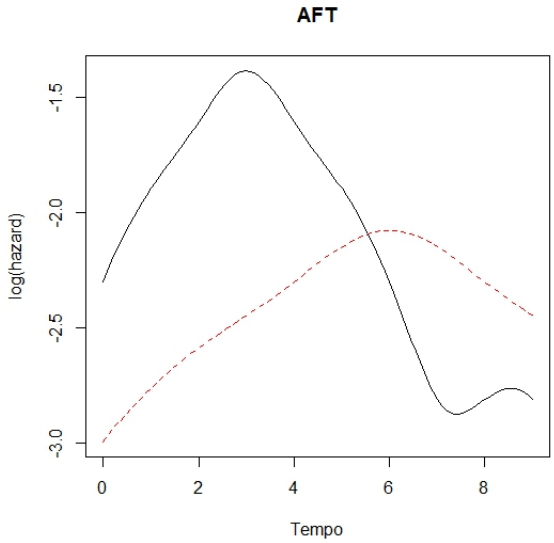

Figura 2.2: Log-hazard in un modello AFT

dove $x_i^T \alpha$ rappresenta la parte parametrica del modello e $\log T_0$ la componente stocastica (Orbe and Nunez-Anton, 2006).

Si osservi che una formulazione alternativa del modello AFT è (Orbe and Nunez-Anton, 2006):

$$\lambda_1(t) = e^{x_i^T \beta} \lambda_0(e^{x_i^T \beta} t) \tag{2.11}$$

Passando ai logaritmi si ottiene:

$$\log \lambda_1(t) = x_i^T \beta + \log \lambda_0(e^{x_i^T \beta} t) \tag{2.12}$$

dove si osserva: $sign(\alpha) = -sign(\beta)$.

In letteratura sono stati proposti altri approcci per la stima e l'inferenza di un modello AFT (Orbe *et al.*, 2002; Orbe and Nunez-Anton, 2006).

Orbe *et al.* (2002) presentano una nuova metodologia, che viene applicata quando non si puó fare l'assunzione di proporzionalitá dei rischi, secondo la quale gli stimatori possono essere ottenuti usando i minimi quadrati pesati; i pesi si ottengono attraverso un'estensione dello stimatore di Kaplan-Meier (Kalbfleisch and Prentice, 2002, 218-244).

Orbe and Nunez-Anton (2006) forniscono un'estensione al modello descritto da Aitkin and Clayton (1980), i quali hanno mostrato come alcuni modelli parametrici per l'analisi della sopravvivenza come, per esempio, i modelli esponenziale o di Weibull, sono facilmente stimati attraverso i modelli lineari generalizzati (GLM). (Per una discussione dettagliata sui GLM si vedano, per esempio McCullagh and Nelder (1988) o Fahrmeir and Tutz (1994)).

Orbe and Nunez-Anton (2006), a differenza di Aitkin and Clayton (1980) utilizzano un GLM semiparametrico, in particolare un modello log-lineare di Poisson.

Nonostante a livello teorico ci siano molti sviluppi sull'argomento, a livello pratico tutti questi approcci sono numericamente complicati e di difficile implementazione, soprattutto quando cresce il numero delle esplicative. Uno dei metodi maggiormente utilizzati è quello di Buckley e James (Buckley and James, 1979) che viene illustrato di seguito.

2.2.1 Metodo di Buckley e James

Uno degli approcci semi-parametrici del modello AFT ampiamente utilizzato e che consente la stima dei parametri della regressione è il metodo di Buckley e James. Buckley and James (1979) introducono un metodo che stima una regressione di minimi quadrati quando sono presenti dati censurati. Il loro stimatore usa le equazioni di stima dei minimi quadrati e un meccanismo di *updating* basato su uno stimatore non parametrico della distribuzione dei residui utile per trattare dati censurati (Jin *et al.*, 2006); (Miller and Halpern, 1982). Il modello assume che il tempo T, o una sua trasformata monotona, sia legata linearmente al vettore di variabili esplicative x, cosicché, indicate con T_i le variabili casuali degli n tempi di sopravvivenza non censurati e con x_i i rispettivi vettori di esplicative, si puó scrivere:

$$T_i = \beta_0 + x_i^T \beta + \epsilon_i \qquad i = 1, ..., n \tag{2.13}$$

dove ϵ_i sono iid con $E[\epsilon_i] = 0$, $Var[\epsilon_i] = \sigma^2$ e distribuzione di ϵ_i indipendente da x.
Dal momento che in presenza di dati censurati si osserva solo

$$Y_i = min(T_i, C_i)$$

dove C_i sono i tempi di censura.
Buckley e James definiscono:

$$Y_i^* = Y_i \delta_i + E(T_i | T_i > Y_i)(1 - \delta_i) \tag{2.14}$$

dove $\delta_i = I(T_i \leq C_i)$. Si dimostra che

$$E[Y_i^*] = E[T_i]. \tag{2.15}$$

è necessario stimare la quantitá $E(T_i|T_i > Y_i)$ per ogni Y_i:

$$\begin{aligned} E(T_i|T_i > Y_i) &= E(\beta_0 + x_i^T\beta + \epsilon_i|(\beta_0 + x_i^T\beta + \epsilon_i) > Y_i) \\ &= \beta_0 + x_i^T\beta + E(\epsilon_i|\epsilon_i > Y_i - (\beta_0 + x_i^T\beta)) \end{aligned} \tag{2.16}$$

e rimane da stimare:

$$\begin{aligned} E(\epsilon_i|\epsilon_i > Y_i - (\beta_0 + x_i^T\beta)) = \\ = \int_{Y_i-(\beta_0+x_i^T\beta)}^{\infty} \frac{\epsilon}{1 - F(Y_i - (\beta_0 + x_i^T\beta))} dF \end{aligned} \tag{2.17}$$

dove F è la funzione di distribuzione di ϵ. Dopo aver sostituito F con una stima $\hat{F}$ di Kaplan-Meier, si ha:

$$y_i^* = y_i\delta_i + \left(x_i^T\tilde{\beta} + \frac{\sum_{\epsilon_j>\epsilon_i} w_j\epsilon_j}{1 - \hat{F}(\epsilon_i)}\right)(1 - \delta_i) \tag{2.18}$$

dove w_j sono i pesi di $\hat{F}$. In questo modo il metodo non dipende da ipotesi sulla distribuzione dei residui. Se si osserva y_i^*, una stima di β potrebbe essere:

$$\hat{\beta} = (X^TX)^{-1}X^Ty^*(\tilde{\beta}) \tag{2.19}$$

dove X è la matrice delle variabili esplicative.
Le formule (2.18) e (2.19) stimano β in maniera iterativa.
Dopo aver stimato β, la stima di β_0 è data da:

$$\hat{\beta_0} = \overline{y}^* - \overline{x}^T\hat{\beta}. \tag{2.20}$$

Questa procedura è molto interessante proprio per l'uso del metodo dei minimi quadrati che consente una facile interpretazione dei risultati e l'utilizzo dell'analisi dei residui, mentre lo schema di *updating* è in generale

difficilmente adattabile alle varie forme di censura. Nel Software R c'è una funzione, *bj* nel pacchetto *rms* che adatta lo stimatore non parametrico dei minimi quadrati di Buckley-James a una variabile risposta censurata a destra. Il programma implementa l'algoritmo come descritto nell'articolo originale di Buckley e James (Buckley and James, 1979). La matrice di varianza e covarianza stimata è basata solo su osservazioni non censurate, ma è stato dimostrato in studi di simulazione che fornisce risultati soddisfacenti. Dal punto di vista computazionale, la convergenza è piuttosto lenta, per cui è necessario aumentare il numero di iterazioni.

2.3 Modello a rischi accelerati

Il modello a rischi accelerati o *accelerated hazard* (AH) è un modello alternativo per dati di sopravvivenza. Rispetto ai modelli di sopravvivenza tipicamente utilizzati, quali per esempio il modello PH e il modello AFT, il modello AH è particolarmente utile per modellare l'effetto di un trattamento o di altre esplicative quando l'effetto del trattamento è graduale e c'è un ritardo (*lag*) prima che il trattamento sia completamente efficace. (Chen *et al.*, 2014).

Il modello (AH) è stato proposto in letteratura da piú di una decina d'anni (Chen and Wang, 2000a; Chen, 2001); tuttavia la sua applicazione è particolarmente limitata dalla complessitá dei metodi di stima semiparametrica esistenti che ostacolano la sua diffusione ed applicazione (Zhang *et al.*, 2011; Chen *et al.*, 2014).

Chen and Wang (2000a) stimano gli effetti della regressione attraverso un'equazione di stima del tipo *non-smooth rank*.

Zhang *et al.* (2011) hanno proposto un nuovo metodo di stima semiparametrica basato su un'approssimazione *kernel-smoothed* della verosimiglianza. (Chen *et al.*, 2014) hanno utilizzato un'approccio Bayesiano non parametrico per la stima del modello.

Chen and Wang (2000a) e Chen (2001) furono i primi a proporre un modello AH che veniva espresso nel seguente modo:

$$\lambda(t, z) = \lambda_0(te^{z^T\gamma}) \tag{2.21}$$

dove $\lambda_0(\cdot)$ è una funzione di rischio *baseline* arbitraria ed ignota e γ è un vettore di parametri ignoti.
L'esponenziale dello i_{-esimo} effetto della regressione, $\exp(\gamma_i)$ è interpretato come un fattore di quanto tempo in piú (o in meno) è richiesto per ottenere lo stesso rischio di fallimento quando lo i_{-esimo} predittore viene incrementato di un'unitá.

Ad esempio nella maggior parte degli studi clinici randomizzati, i soggetti sono casualmente assegnati al gruppo dei trattati e al gruppo dei controlli prima che essi ricevano il trattamento o il placebo. Per valutare l'efficacia dei differenti trattamenti terapeutici, assumiamo che $\lambda_1(t)$ sia la funzione *hazard* del gruppo dei trattati e $\lambda_0(t)$ sia la funzione *hazard* del gruppo dei controlli. Se la randomizzazione è ben bilanciata, è piú ragionevole assumere che le funzioni *hazard* non siano distinguibili a $t = 0$. In questo contesto, l'utilizzo dei modelli PH ed AFT sarebbe errato in quanto entrambi i modelli impongono che le funzioni *hazard* siano identiche per tutti $t \geq 0$ se $\lambda_1(0) = \lambda_0(0)$. Il modello AH, di contro, è piú indicato per questo obiettivo, perché garantisce, anche se $\gamma \neq 0$, che $\lambda_1(0) = \lambda_0(0)$ senza forzare le intere funzioni ad essere identiche (Zhang *et al.*, 2011). Inoltre,

il modello AH puó usare anche un solo parametro nelle funzioni *hazard* per caratterizzare un fenomeno, mentre i modelli PH o AFT dovrebbero introdurre interazioni o termini dipendenti dal tempo per modellare eventuali variazioni.
Il beneficio dell'effetto di un trattamento nel modello AH è ben definito quando le funzioni *hazard* sono monotone, poiché l'ordine viene ancora mantenuto sia nell'accelerazione sia nella decelerazione. Un'accelerazione di $\gamma > 0$ significa che l'effetto del trattamento è benefico (dannoso) se le funzioni *hazard* sono monotone decrescenti (crescenti). Di contro un'accelerazione di $\gamma < 0$ implica che l'effetto del trattamento è benefico (dannoso) se le funzioni *hazard* sono monotone crescenti (decrescenti).

In riferimento alla funzione *hazard* del gruppo di controllo, il trattamento in un modello AH non modifica la forma complessiva della funzione di rischio, ma piuttosto riscala il suo asse temporale (Chen and Wang, 2000a). Cioé:

$$\lambda_1(t) = \lambda_0(te^{\gamma}) \tag{2.22}$$

dove e^{γ} indica come il trattamento abbia l'effetto di far variare la scala temporale della sottostante funzione *hazard* ed è, pertanto, chiamato *hazard progression time ratio*. Il valore di γ riflette la forza e la direzione di tale cambiamento di scala. La direzione dell'alterazione puó essere o di accelerazione o di decelerazione, dipende se $\gamma > 0$ o $\gamma < 0$, rispettivamente.
Per esempio, se $\gamma = -\log 2$, il rischio per gli individui appartenenti al gruppo dei trattati progredisce in metá tempo rispetto al rischio per gli individui appartenenti al gruppo di controllo. Se, invece, $\gamma = \log 2$, il rischio per gli individui appartenenti al gruppo dei trattati progredisce nel doppio del tempo rispetto al rischio per gli individui appartenenti al gruppo di controllo.

Ovviamente, non c'è nessuna differenza tra i due gruppi se $\gamma = 0$ (Chen, 2001).

La Figura 2.3 mostra a livello esemplificativo il logaritmo dell'hazard calcolato per due specifici gruppi di soggetti (gruppo 0 e gruppo 1), in un modello AH.
Si osservi che per $t = 0$ le due funzioni di rischio sono uguali.

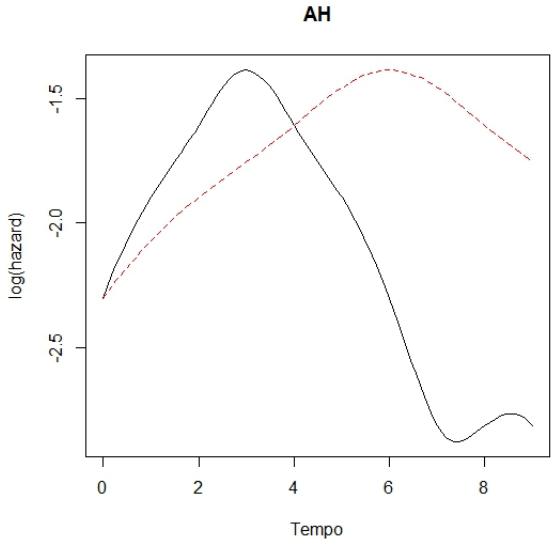

Figura 2.3: Log-hazard in un modello AH

Poiché i modelli PH, AFT ed AH caratterizzano differenti aspetti del trattamento, tali effetti non sempre sono confrontabili. È difficile comprendere quale modello sia piú vicino alla realtá, dipende essenzialmente dalle

sue assunzioni e da quanto queste siano vicine alla realtá. Per esempio, quando ci si aspetta che il trattamento sia altamente efficace e il suo effetto sia persistente, sono preferibili i modelli PH o AFT; quando ci si aspetta che il trattamento in uno studio bilanciato abbia un effetto graduale, è preferibile il modello AH.

Nel capitolo successivo viene proposta una classe generale di modelli di regressione semiparametrica per la funzione *hazard*. Questa classe generale include i modelli PH, AFT ed AH.

Capitolo 3

Un Modello di Regressione Esteso

In questo capitolo viene proposta una classe generale di modelli di regressione semiparametrica per la modellazione della funzione *hazard*. Questa classe generale include come sottoclassi le classi di modelli popolari viste in precedenza, come il modello PH di Cox, il modello AFT e il modello AH. Il nuovo modello è flessibile nell'adattarsi ai dati di sopravvivenza e consente di predire il processo di sopravvivenza di un soggetto. Grazie alla sua struttura che include il modello PH, il modello AFT e il modello AH, questa classe generale di modelli puó fornire uno strumento per determinare quale di essi è piú appropriato per i dati in esame (Chen and Jewell, 2001).

In letteratura sono presenti diversi articoli dove vengono proposti modelli estesi che possano essere utilizzati nei molteplici scenari che gli studi di sopravvivenza possono presentare (Etezadi-Amoli and Ciampi, 1987;

Tseng *et al.*, 2014; Tseng and Shu, 2011; Tong *et al.*, 2013).
Etezadi-Amoli and Ciampi (1987) discutono un modello esteso che include sia il modello a rischi proporzionali di Cox sia il modello AFT. Gli autori utilizzano le spline per approssimare la funzione hazard baseline e sviluppano una procedura di stima basata sulla massima verosimiglianza per fornire le stime della funzione hazard e dei coefficienti di regressione.
Tseng *et al.* (2014) introducono una classe generale di modelli semiparametrici, chiamati *extended hazard* con lo scopo di essere applicati ai diversi contesti in cui i dati di sopravvivenza con esplicative dipendenti dal tempo possono essere utilizzati. Questo modello esteso contiene come sottoclassi sia il modello a rischi proporzionali di Cox sia il modello AFT. Per la stima dei parametri del modello proposto, gli autori sviluppano una classe di equazioni di stima che utilizza processi di conteggio e tecniche Martingale.
Tseng and Shu (2011) propongono un modello alternativo per la stima dei modelli *extended hazard* (Tseng *et al.*, 2014), basato su un'approssimazione *kernel-smoothed* della verosimiglianza. Tale modello esteso contiene come sottoclassi sia il modello a rischi proporzionali di Cox sia il modello AFT.
Tong *et al.* (2013) come Tseng and Shu (2011) si basano su un'approssimazione *kernel-smoothed* della verosimiglianza per la stima di un modello semiparametrico esteso che contiene come sottoclassi i modelli PH ed AFT. Inoltre, Tong *et al.* (2013) sviluppano una tecnica di selezione con struttura penalizzata per determinare quali esplicative costituiscono il modello PH e quali il modello AFT.

La classe generale di modelli di regressione semiparametrica per la funzione *hazard* che in questa tesi si propone, include i modelli PH, AFT ed

anche i modelli AH, che nei modelli estesi presenti in letteratura, in genere, non vengono considerati.

Si denoti con $\lambda(t, x, z)$ la funzione *hazard* al tempo t per un soggetto con due vettori di esplicative x e z. Possiamo definire il modello di regressione *Extended Hazard* (EH) nel seguente modo:

$$\lambda(t, x, z) = \lambda_0(te^{z^T\gamma})\, e^{x^T\beta}. \tag{3.1}$$

È facile vedere che la formula si riduce al modello PH per $\gamma = 0$, al modello AFT per $\gamma = \beta$ quando $x \equiv z$ e, infine, al modello AH per $\beta = 0$.

La Tabella 3.1 mostra che i modelli PH, AFT ed AH sono annidati in questo modello esteso.

Tabella 3.1: Modello di Regressione Esteso

Condizione	*Modello*	$\lambda(t, x, z)$
$\gamma = 0$	PH	$\lambda_0(t)\, e^{x^T\beta}$
$\gamma = \beta$ ($se\ x \equiv z$)	AFT	$\lambda_0(te^{x^T\beta})\, e^{x^T\beta}$
$\beta = 0$	AH	$\lambda_0(te^{z^T\gamma})$

Come esempio assumiamo che l'esplicativa X sia binaria.

Supponiamo che $X = 1$ si riferisca al gruppo dei trattati e $X = 0$ al gruppo dei controlli, come in uno studio clinico randomizzato. Allora il modello si riduce a:

$$\lambda_1(t) = \lambda_0(te^{\gamma})\, e^{\beta}, \tag{3.2}$$

dove $\lambda_1(\cdot)$ è la funzione *hazard* del gruppo dei trattati, $\lambda_0(\cdot)$ è la funzione *hazard* del gruppo dei controlli, e γ e β sono i parametri. I due parametri, γ e β, possono essere interpretati come misure di due differenti effetti che l'esplicativa puó avere sulla sopravvivenza. Il primo parametro, γ, identifica l'accelerazione/decelerazione della progressione del rischio nel gruppo dei trattati, mentre β caratterizza l'*hazard* relativo dopo aver corretto per le differenti progressioni del rischio nei gruppi dei trattati e dei controlli. Quindi, il modello EH implica che il trattamento puó modificare simultaneamente la grandezza del rischio e la velocitá della progressione.

Identificare e stimare correttamente queste due componenti puó consentire una migliore descrizione dei dati e portare ad una piú accurata modellazione della sopravvivenza dei soggetti in studio, sebbene il principale valore di questo modello generale è quello di quantificare e discriminare tra le tre sottoclassi di modelli che di solito vengono usate individualmente.

La Figura 3.1 mostra a livello esemplificativo il logaritmo dell'hazard calcolato per due specifici gruppi di soggetti (gruppo 0 e gruppo 1), nei diversi modelli.

Si osservi che il grafico relativo al modello esteso *EH* rappresenta una generalizzazione dei tre grafici visti: *PH*, dove si osserva la proporzionalitá dei rischi evidenziata dalle curve parallele; *AH*, dove per $t = 0$ le due funzioni di rischio sono uguali; e *AFT*, dove si ha l'uguaglianza tra i due gruppi di esplicative ($x \equiv z$).

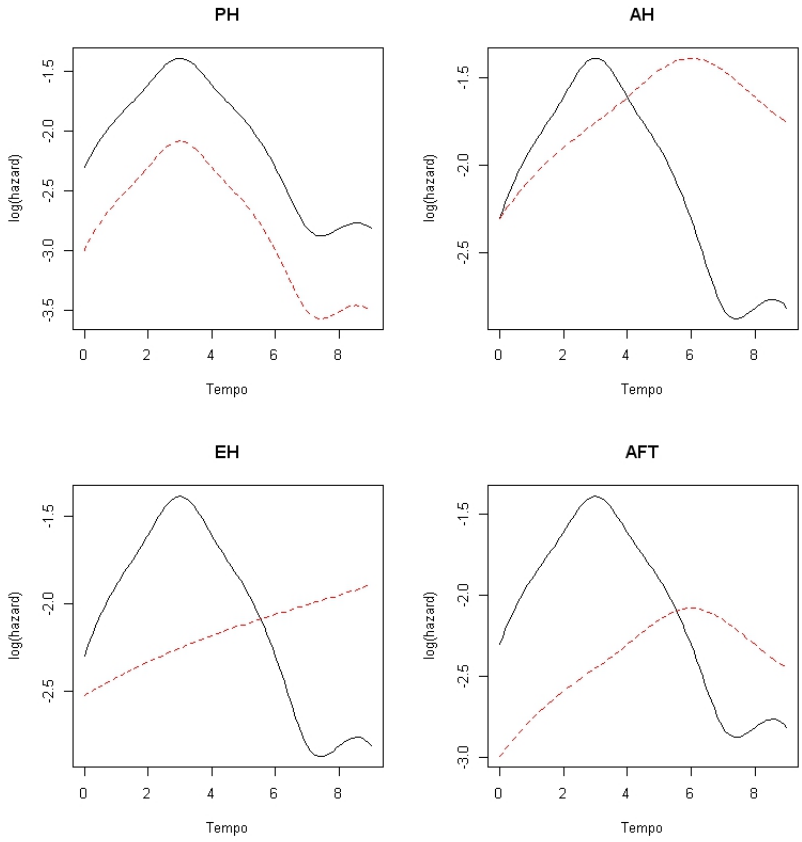

Figura 3.1: Log-hazard nei diversi modelli

Per stimare il modello (3.1) è necessario introdurre una formulazione alternativa dei tempi di sopravvivenza, pertanto, nel paragrafo successivo viene introdotto il modello di regressione di Poisson.

3.1 Modello di Regressione di Poisson

Se gli studi di sopravvivenza vengono analizzati alla luce della tradizione demografica, l'osservazione alla base dello studio stesso non è il tempo del verificarsi del evento (o della censura) per ciascun soggetto dello studio, ma piuttosto, molti piccoli intervalli di follow-up per ciascun soggetto (Carstensen, 2005; Whitehead, 1980). L'obiettivo dell'analisi diventa, pertanto, quello di modellare i tassi piuttosto che il tempo del verificarsi dell'evento; la variabile risposta è, quindi, un *outcome* di tipo 0/1 in ciascun intervallo. Il tempo viene visto come una variabile esplicativa piuttosto che come una risposta.

In questo contesto uno strumento molto utile è il modello di regressione di Poisson. L'implementazione di un modello di Poisson in pratica richiede che il *follow-up* per ciascun soggetto sia suddiviso in piccoli intervalli di *follow-up*. Le esplicative tempo dipendenti verranno calcolate per ciascun intervallo, mentre quelle che hanno valori fissi nel tempo verranno trascritte, per ciascun soggetto, in tutti gli intervalli.

A livello pratico, l'implementazione di un modello di Poisson richiede l'utilizzo di tante osservazioni per ciascun soggetto e questo in passato poteva essere un problema da un punto di vista computazionale a causa della mancanza di *software* adatti ad un tale approccio. Adesso, invece sono disponibili diverse soluzioni con Stata, SAS ed R che rendono tale approccio ampiamente accessibile.

Il vantaggio che rimane del modello di Cox è la sua abilitá a produrre facilmente le stime delle probabilitá di sopravvivenza negli studi clinici con un ben definito tempo di inizio per tutti gli individui. Questo puó, tuttavia, essere realizzato anche attraverso un modello di Poisson con dei metodi che non risultano eccessivamente complicati.

3.2 La variabile Tempo: Risposta o Esplicativa?

Generalmente, l'analisi della sopravvivenza si basa sullo studio di dati del tipo (T, C), dove si osserva soltanto $min(T, C)$ e $\delta = 1\{T < C\}$. Questo è un approccio che considera il tempo di sopravvivenza T come variabile risposta, anche se non completamente osservata perché limitata dal tempo di censura C.

Si consideri uno studio (di sopravvivenza) di *follow-up* dove il tempo di *follow-up* per ciascun soggetto viene suddiviso in piccoli intervalli di uguale lunghezza y. Per ciascun intervallo si rilevi la variabile d'interesse che assumerá valore 0 per tutti gli intervalli e valore 1 soltanto per l'ultimo intervallo nel quale il soggetto sperimenta l'evento in studio (Carstensen, 2005). Ciascun piccolo intervallo per ogni soggetto fornisce un'osservazione di quello che puó essere definito *tasso empirico*, (d, y), dove d è il numero di eventi nell'intervallo, che tipicamente è (0 o 1), e y è la lunghezza dell'intervallo, cioé quello che viene definito *risk time*.

Il tasso teorico di occorrenza dell'evento è definito come il tasso:

$$\lambda(t) = \lim_{h \to 0} \frac{P\{evento\ in\ (t, t+h) \mid a\ rischio\ al\ tempo\ t\}}{h} \tag{3.3}$$

che puó dipendere da un certo numero di esplicative. Con questa formulazione il tempo (scala) t ha lo *status* di un'esplicativa e il *risk time* h, che è la differenza tra due punti nella scala temporale, lo *status* proprio di *risk time*.

3.2.1 Verosimiglianza per tassi empirici

In un intervallo di lunghezza y con tasso costante λ, la probabilitá dell'evento è λy, per cui il contributo alla verosimiglianza di ciascun tasso empirico

(d, y) è la verosimiglianza di Bernoulli che è cosí definita:

$$L(\lambda|(d,y)) = (\lambda y)^d \times (1-\lambda y)^{1-d} = \left(\frac{\lambda y}{1-\lambda y}\right)^d (1-\lambda y) \tag{3.4}$$

Passando ai logaritmi si ha:

$$\ell(\lambda|(d,y)) = d\ \log\left(\frac{\lambda y}{1-\lambda y}\right) + \log(1-\lambda y) \approx d\ \log(\lambda) + d\ \log(y) - \lambda y \tag{3.5}$$

Questa formulazione si ottiene attraverso l'approssimazione di Poisson alla Binomiale.

L'osservazione di tassi empirici indipendenti con lo stesso parametro da luogo ad una log-verosimiglianza che dipende dai tassi empirici solo attraverso $D = \sum d$ e $Y = \sum y$, ovvero:

$$\ell(\lambda|(D,Y)) = D\ \log(\lambda) - \lambda Y \tag{3.6}$$

che è la log-verosimiglianza di una variabile Poisson D con media λY.

Si osservi che i contributi alla verosimiglianza di ciascun soggetto non saranno indipendenti, ma saranno condizionatamente indipendenti; la verosimiglianza totale di ciascun individuo sará il prodotto delle probabilitá condizionate della forma:

$$\begin{aligned} P\{\textit{evento in } (t_3,t_4) \mid \textit{vivo a } t_3\} \ &\times\ P\{\textit{sopravvivenza } (t_2,t_3) \mid \textit{vivo a } t_2\} \\ &\times\ P\{\textit{sopravvivenza } (t_1,t_2) \mid \textit{vivo a } t_1\} \\ &\times\ P\{\textit{sopravvivenza } (t_0,t_1) \mid \textit{vivo a } t_0\} \end{aligned} \tag{3.7}$$

Quindi la verosimiglianza per un insieme di tassi empirici si presenta come una verosimiglianza per osservazioni di Poisson indipendenti, ma

non lo è in quanto è un prodotto (e la log-verosimiglianza una somma) di probabilitá condizionate. Cosí gli studi di sopravvivenza possono essere analizzati usando la verosimiglianza di Poisson per osservazioni indipendenti; dipende da quanto si è disposti ad accettare in termini di larghezza degli intervalli. Si noti che è soltanto la verosimiglianza che coincide con quella di un modello di Poisson, non la distribuzione della variabile risposta (d, y), quindi, è ammissibile solo l'inferenza basata sulla verosimiglianza.
L'analisi di un modello moltiplicativo per il parametro λ permette di adattare un modello di Poisson:

$$log(\mu) = log(\lambda Y) = log(\lambda) + log(Y). \tag{3.8}$$

Se $log(\lambda)$ è il parametro, il termine $log(Y)$ deve essere considerato un'esplicativa con coefficiente uguale a 1, il cosiddetto *offset* (McCullagh and Nelder, 1988).

3.3 Il modello di Cox attraverso la formulazione di Poisson

In questo paragrafo viene descritto il modo attraverso cui un modello di Cox viene stimato utilizzando un modello di Poisson.
Si consideri il modello di Cox nella sua formulazione classica:

$$\lambda(t, x) = \lambda_0(t) \times \exp(\eta), \quad \eta = X\beta \tag{3.9}$$

Il modello di Cox considera la log-verosimiglianza parziale per i parametri $\beta = (\beta_1, \cdots, \beta_p)$ nel predittore lineare $\eta_i = \beta_1 x_{1i} + \cdots + \beta_p x_{pi}$:

$$\ell(\beta) = \sum_{eventi} \log\left(\frac{e^{\eta_{evento}}}{\sum_{i \in \Re_t} e^{\eta_i}}\right) \tag{3.10}$$

dove $\mathfrak{R}_t$ è l'insieme dei soggetti a rischio al tempo t.
Supponiamo che la scala temporale sia suddivisa in piccoli intervalli di tempo con al piú un evento in ciascuno di essi e che in aggiunta ai parametri della regressione che descrivono l'effetto delle esplicative, si usi un parametro per descrivere l'effetto del tempo. Quindi, il modello con tassi costanti in ciascun piccolo intervallo è:

$$\log(\lambda(t, x_i)) = \log(\lambda_0(t)) + \beta_1 x_{1i} + ... + \beta_p x_{pi} = \alpha_t + \eta_i. \tag{3.11}$$

dove $\alpha_t = f(t)$, ovvero α_t viene stimato attraverso le spline o i polinomi.
I contributi alla log-verosimiglianza che contengono informazioni su uno specifico parametro della scala temporale α_t, saranno contributi dall'*empirical rate* $(d, y) = (1, 1)$ con l'evento al tempo t, e tutti i tassi empirici $(d, y) = (0, 1)$ da tutti gli altri individui che sono a rischio al tempo t. C'è esattamente un contributo da ciascun soggetto a rischio alla log-verosimiglianza:

$$\ell_t(\alpha_t, \beta) = \sum_{i \in \mathfrak{R}_t} \{d_i(\alpha_t + \eta_i) - e^{\alpha_t + \eta_i}\} = \alpha_t + \eta_{evento} - e^{\alpha_t} \sum_{i \in \mathfrak{R}_t} e^{\eta_i} \tag{3.12}$$

dove η_{evento} è il predittore lineare per il soggetto che sperimenta l'evento. La derivata rispetto a α_t è:

$$\frac{\partial \ell_{(\alpha_t, \beta)}}{\partial_{\alpha_t}} = 1 - e^{\alpha_t} \sum_{i \in \mathfrak{R}_t} e^{\eta_i} = 0 \iff e^{\alpha_t} = \frac{1}{\sum_{i \in \mathfrak{R}_t} e^{\eta_i}} \tag{3.13}$$

Se la stima di e^{α_t} viene inserita nella log-verosimiglianza per α_t, otteniamo la *verosimiglianza profilo*:

$$\log\left(\frac{1}{\sum_{i \in \mathfrak{R}_t} e^{\eta_i}}\right) + \eta_{death} - 1 = \log\left(\frac{e^{\eta_{death}}}{\sum_{i \in \mathfrak{R}_t} e^{\eta_i}}\right) - 1 \tag{3.14}$$

che è la stessa del contributo al tempo t in una log-verosimiglianza parziale di Cox.

Il modello di Cox potrebbe, quindi, essere stato formulato come un modello dove c'è un parametro della scala temporale separato per ciascun intervallo di tempo. Per quegli intervalli sulla scala temporale dove non si è verificato l'evento in studio, la stima di α_t sará $-\infty$, e cosí questi intervalli non contribuiranno alla log-verosimiglianza.

Quindi, un modello di Cox puó essere stimato attraverso un modello di regressione di Poisson *esplodendo* opportunamente i dati e specificando un modello che abbia un parametro per ogni intervallo temporale. I risultati saranno analoghi anche in termini di errori standard, poiché è la stessa la verosimiglianza che viene massimizzata.

Questa non è una nuova scoperta, giá Whitehead (1980) aveva evidenziato questa possibilitá. Tuttavia, i problemi a livello computazionale connessi con questo approccio erano eccessivi per consentire al metodo di diffondersi a livello pratico in quegli anni.

Riassumendo, possiamo asserire che l'implementazione di un modello di Poisson in pratica richiede che il tempo di sopravvivenza per ciascun soggetto sia suddiviso in piccoli intervalli di *follow-up*. Le esplicative che variano nel tempo verranno calcolate per ciascun intervallo, mentre quelle che hanno valori fissi nel tempo verranno ripetute, per ciascun soggetto, in tutti gli intervalli. Quindi, si *esplode* un *dataset* tradizionale per dati di sopravvivenza con un *record* per ogni soggetto in un *dataset* con diversi *record* per ogni soggetto, uno per ogni intervallo di *follow-up*.
In un modello di Poisson il logaritmo del *risk-time* viene inserito come *offset* e il tempo come esplicativa. Quindi, un modello di Poisson per dati di

follow-up consente di fare una netta distinzione tra *risk-time* come variabile risposta e *time-scale* come esplicativa. Dopo che i dati vengono suddivisi in piccoli intervalli di *follow-up*, l'effetto di ciascun *time-scale* puó essere stimato usando le *spline* o i *polinomi*. Cosí si otterranno direttamente i tassi *baseline* stimati attraverso un modello lineare generalizzato con un dato insieme di esplicative.

Capitolo 4

Stima del modello di regressione esteso

Nel presente capitolo viene proposta una procedura di stima del modello di regressione esteso per l'analisi semiparametrica dei dati di sopravvivenza, sfruttando la formulazione di Poisson descritta nel capitolo precedente per il modello di Cox. L'obiettivo, quindi è quello di utilizzare questa formulazione alternativa che fa uso del modello di regressione di Poisson, per stimare i modelli visti, modello di Cox a rischi proporzionali (PH), modello *Accelerated Failure Time* (AFT) e, infine, modello *Accelerated Hazard* (AH), che in letteratura sono stimati attraverso l'ottimizzazione di criteri molto diversi tra loro.

Si consideri il modello generale visto nel precedente capitolo (3.1), dove, per esemplificare si suppongono x e z non vettoriali:

$$\lambda(t, x, z) = \lambda_0(te^{z\gamma})\, e^{x\beta}. \tag{4.1}$$

Come discusso, il modello include come casi particolari:

- Se $\gamma = 0 \implies$ *modello di Cox PH* $\quad \lambda(t, x) = \lambda_0(t)\, e^{x\beta}$

- Se $\gamma = \beta\ (x = z) \implies$ *modello AFT* $\quad \lambda(t, x) = \lambda_0(te^{x\beta})\, e^{x\beta}$

- Se $\beta = 0 \implies$ *modello AH* $\quad \lambda(t, z) = \lambda_0(te^{z\gamma})$

Come abbiamo evidenziato nel precedente capitolo, i due parametri, γ e β, possono essere interpretati come misure di due differenti effetti che le esplicative possono avere sulla sopravvivenza.
Il primo parametro, γ, identifica l'accelerazione/decelerazione della progressione del rischio nei diversi gruppi di soggetti (gruppi che si costituiscono in relazione all'esplicativa in esame); mentre β caratterizza l'*hazard* relativo dopo aver corretto per le differenti progressioni del rischio nei differenti gruppi.
Il modello esteso implica, quindi, che l'esplicativa possa modificare simultaneamente la grandezza del rischio e la velocitá della progressione.
Stimare in maniera accurata queste due componenti puó consentire una migliore descrizione dei dati. Infatti, un valore di questo modello generale per l'analisi dei dati di sopravvivenza potrebbe essere quello di avere una maggiore flessibilitá nell'adattamento oltre a consentire confronti tra i modelli PH, AFT, ed AH, modelli tra di loro differenti e, di conseguenza, non immediatamente confrontabili.

4.1 Algoritmo di stima

L'equazione (4.1) esprime una forma funzionale per il parametro di una variabile casuale Poisson, dove $\lambda_0(\cdot)$ viene espressa attraverso polinomi o basi di spline. Utilizzando una formulazione log-lineare, nello spirito di una classica regressione di Poisson si ottiene:

$$\log(\lambda(t, x, z)) = \log(\lambda_0(te^{z\gamma})) + x\beta \tag{4.2}$$

che definisce una funzione non lineare nei parametri. La diretta massimizzazione della log-verosimiglianza di Poisson potrebbe essere utilizzata per ottenere le stime dei parametri, tuttavia suggeriamo in questo capitolo una procedura che consente di stimare il modello (4.2) attraverso modelli *lineari* di Poisson.

Il modello (4.2) contiene una funzione che è non lineare in γ, pertanto, il modello viene riscritto considerando una funzione $f(\cdot)$ liscia e non specificata, tale che:

$$\log(\lambda_0(te^{z\gamma})) = f(\gamma, t). \tag{4.3}$$

Per linearizzare tale funzione, si consideri lo sviluppo in serie di Taylor:

$$f(\gamma, t) \approx f(\tilde{\gamma}, t) + (\gamma - \tilde{\gamma})\, f'(\tilde{\gamma}, t) \tag{4.4}$$

dove si assume $\tilde{\gamma}$ noto.
Considerando che nel modello in questione la funzione è:

$$f(te^{z\gamma}),$$

si ottiene:

$$f(te^{z\gamma}) \approx f(te^{z\tilde{\gamma}}) + (\gamma - \tilde{\gamma})\, f'(te^{z\tilde{\gamma}})\; t\; e^{z\tilde{\gamma}} z, \tag{4.5}$$

Ponendo $\widetilde{W} = f'(te^{z\tilde{\gamma}})\, t\, e^{z\tilde{\gamma}} z$, il modello (4.2) puó essere formulato come segue:

$$\log(\lambda(t, x, z)) = f(te^{z\tilde{\gamma}}) + \gamma\widetilde{W} - \tilde{\gamma}\widetilde{W} + x\beta \tag{4.6}$$

ovvero:

$$\log(\lambda(t, x, z)) = \gamma\widetilde{W} + x\beta + f(te^{z\tilde{\gamma}}) - \tilde{\gamma}\widetilde{W} \tag{4.7}$$

dove $\tilde{\gamma}\widetilde{W}$ è l'*offset*, $f(te^{z\tilde{\gamma}})$ puó essere stimato attraverso le *spline* o i polinomi e, infine, β e γ sono coefficienti lineari.

L'equazione di regressione (4.7) definisce un'approssimazione lineare della (4.2), assumendo un valore $\tilde{\gamma}$ noto. Questo suggerisce di stimare il modello di Poisson (4.2) attraverso i seguenti passi.

1. Si fissi un valore iniziale $\tilde{\gamma}$
2. Si calcoli la *pseudo* variabile esplicativa

$$\widetilde{W} = f'(te^{z\tilde{\gamma}})\, t\, e^{z\tilde{\gamma}} z$$

3. Si stimi una regressione lineare di Poisson con equazione

$$\log(\lambda(t, x, z)) = \gamma\widetilde{W} + x\beta + f(te^{z\tilde{\gamma}}) + Offs$$

   dove $Offs = -\tilde{\gamma}\widetilde{W}$ è il termine *offset*, ovvero una variabile con coefficiente noto e fissato uguale a 1

4. Ottenute le stime $(\hat{\gamma}, \hat{\beta}, \hat{f})$, si ponga $\hat{\gamma} \longrightarrow \tilde{\gamma}$
5. Si ripetano i passi 2 - 4 fino a convergenza.

L'algoritmo descritto sembra funzionare in modo adeguato nella pratica. Nelle analisi e negli studi di simulazione riportati nei capitoli successivi, è stato impiegato $\tilde{\gamma} = 0$ che sembra funzionare bene. Inoltre, è stato osservato che una *correzione del passo* garantisce una maggiore stabilitá con un aumento non rilevante nel numero delle iterazioni. Infine, come criterio di arresto per definire la convergenza sono state considerate variazioni relative della devianza.

L'algoritmo descritto si riferisce al modello EH. Tuttavia è possibile impiegarlo anche per stimare i casi particolari: PH, AFT ed AH.
Si consideri l'equazione (4.7) nei tre distinti modelli. In particolare, se si vuole stimare il modello PH di Cox non è necessario iterare poiché tale modello si ottiene a partire dal modello (4.7) ponendo $\gamma = 0$. Per cui si ha:

$$\log(\lambda(t, x)) = f(t) + x\beta \tag{4.8}$$

Se si vuole stimare il modello AFT bisogna considerare $\gamma = \beta$ e $x = z$, per cui la stima del modello si ha attraverso i seguenti passi:

1. Si fissi un valore iniziale $\tilde{\beta}$
2. Si calcoli la *pseudo* variabile esplicativa

$$\widetilde{W} = f'(te^{x\tilde{\beta}})\, t\, e^{x\tilde{\beta}} x$$

3. Si stimi una regressione lineare di Poisson con equazione

$$\begin{aligned} \log(\lambda(t, x)) &= \beta\widetilde{W} + x\beta + f(te^{x\tilde{\beta}}) - \tilde{\beta}\widetilde{W} \\ &= \beta(\widetilde{W} + x) + f(te^{x\tilde{\beta}}) - \tilde{\beta}\widetilde{W} \end{aligned} \tag{4.9}$$

4. Ottenute le stime $(\hat{\beta}, \hat{f})$, si ponga $\hat{\beta} \longrightarrow \tilde{\beta}$

5. Si ripetano i passi 2 - 4 fino a convergenza.

Se si vuole stimare il modello AH bisogna considerare $\beta = 0$, per cui la stima del modello si ottiene con il seguente algoritmo:

1. Si fissi un valore iniziale $\tilde{\gamma}$

2. Si calcoli la *pseudo* variabile esplicativa

$$\widetilde{W} = f'(te^{z\tilde{\gamma}})\, t\, e^{z\tilde{\gamma}} z$$

3. Si stimi una regressione lineare di Poisson con equazione

$$\log(\lambda(t, z)) = \gamma \widetilde{W} + f(te^{z\tilde{\gamma}}) - \tilde{\gamma}\widetilde{W} \tag{4.10}$$

4. Ottenute le stime $(\hat{\gamma}, \hat{f})$, si ponga $\hat{\gamma} \longrightarrow \tilde{\gamma}$

5. Si ripetano i passi 2 - 4 fino a convergenza.

Capitolo 5

Simulazioni

Nel presente capitolo vengono illustrati i risultati ottenuti attraverso delle simulazioni di tipo Montecarlo. Tali simulazioni sono state svolte al fine di verificare l'attendibilitá dell'algoritmo suggerito e descritto nel capitolo precedente per la stima dei dati di sopravvivenza attraverso il modello di regressione di Poisson.
Per poter procedere con le simulazioni sono stati considerati i seguenti scenari.
Sia η il predittore lineare:

$$\eta = 2 + \beta_1 X_1 + \beta_2 X_2$$

dove $X_1 \sim Ber(0.5),\ X_2 \sim N(0, 0.5)$.

Si consideri, in particolare:

$$\beta_1 = \beta_2 = 1.$$

Si considerino diverse distribuzioni per il tempo di sopravvivenza T. Si osservi, come descritto nella Tabella 1.2, che le distribuzioni Gumbel, Normale e Logistica per ϵ corrispondono a distribuzioni Weibull, log-Normale

e log-Logistica per T.
Pertanto si consideri:

$$T \sim Weibull \Leftrightarrow \log(T) \sim Gumbel$$

$$T \sim logNormale \Leftrightarrow \log(T) \sim Normale$$

$$T \sim logLogistica \Leftrightarrow \log(T) \sim Logistica$$

I tempi di sopravvivenza sono stati simulati attraverso:

$$T \sim \exp(\eta + \sigma * \epsilon)$$

dove ϵ sono realizzazioni da una Gumbel, Normale o Logistica, η è il predittore lineare e σ è il parametro di scala.

Per avviare le simulazioni sono stati considerati, altresí, il numero di repliche, la numerositá campionaria e la percentuale di censura come segue:

- il numero di repliche è posto pari a 300;
- la numerositá campionaria è $n = 100, 200, 500$;
- la percentuale di censura è considerata pari al 30%.

Infine, bisogna sottolineare che le simulazioni sono state effettuate per due differenti valori del parametro di scala:

$$\sigma = 0.5, 1.$$

L'implementazione dell'algoritmo proposto per il modello PH di Cox è banale e non viene ulteriormente discusso; quello per il modello AH risulta difficilmente confrontabile in quanto, come discusso nel paragrafo 2.3, la stima di tale modello risulta complicata e non esistono al momento librerie o pacchetti R disponibili. Quindi le simulazioni effettuate riguardano solo il modello AFT.
Per i diversi scenari descritti sopra, confrontiamo gli stimatori di β_1 e β_2 ottenuti attraverso l'algoritmo proposto per il modello AFT, il metodo di Buckley e James descritto nel paragrafo 2.2.1 e implementato attraverso la funzione bj(rms) e, infine, il metodo parametrico, implementato attraverso la funzione survreg(survival) che puó essere considerato il *gold standard*.
Per poter comprendere se le stime ottenute sono *vicine* ai veri valori dei parametri sono stati riportati nelle tabelle seguenti le medie e le deviazioni standard dei parametri stimati con i diversi modelli. Vengono riportate, altresí le mediane per verificare l'eventuale presenza di dati anomali.

Tabella 5.1: Risultati per distribuzione GUMBEL con $\sigma = 0.5$

			Distribuzione campionaria		
n		metodo	media	mediana	dev. std
100	β_1	AFT	1,000	1,002	0,117
		BJ	0,991	0,992	0,147
		PAR	0,996	0,996	0,118
	β_2	AFT	1,014	1,017	0,110
		BJ	1,001	0,997	0,132
		PAR	1,003	1,005	0,108
200	β_1	AFT	1,010	1,011	0,085
		BJ	1,005	1,000	0,103
		PAR	1,007	1,006	0,084
	β_2	AFT	1,004	1,001	0,095
		BJ	0,998	0,994	0,110
		PAR	0,997	0,996	0,093
500	β_1	AFT	1,009	1,010	0,054
		BJ	1,006	1,004	0,066
		PAR	1,002	1,003	0,053
	β_2	AFT	1,005	1,003	0,053
		BJ	1,000	1,000	0,063
		PAR	1,000	0,999	0,052

Tabella 5.2: Risultati per distribuzione GUMBEL con $\sigma = 1$

			Distribuzione campionaria		
n		metodo	media	mediana	dev. std
100	β_1	AFT	0,991	0,988	0,243
		BJ	0,991	0,978	0,303
		PAR	0,995	0,991	0,242
	β_2	AFT	1,000	0,998	0,207
		BJ	1,005	1,025	0,249
		PAR	1,002	1,002	0,207
200	β_1	AFT	0,933	1,002	0,170
		BJ	1,010	1,005	0,211
		PAR	1,007	1,003	0,176
	β_2	AFT	1,075	0,999	0,191
		BJ	0,993	0,996	0,217
		PAR	0,993	1,001	0,182
500	β_1	AFT	0,992	0,989	0,102
		BJ	0,989	0,986	0,129
		PAR	0,993	0,992	0,103
	β_2	AFT	1,001	1,003	0,099
		BJ	1,002	1,005	0,124
		PAR	1,001	1,004	0,100

Tabella 5.3: Risultati per distribuzione NORMALE con $\sigma = 0.5$

			Distribuzione campionaria		
n		metodo	media	mediana	dev. std
100	β_1	AFT	1,024	1,024	0,139
		BJ	1,008	1,011	0,121
		PAR	1,008	1,008	0,121
	β_2	AFT	1,033	1,031	0,129
		BJ	0,999	1,000	0,109
		PAR	0,999	1,001	0,108
200	β_1	AFT	1,024	1,024	0,099
		BJ	1,003	1,006	0,088
		PAR	1,004	1,007	0,088
	β_2	AFT	1,038	1,043	0,100
		BJ	1,005	1,007	0,086
		PAR	1,006	1,008	0,087
500	β_1	AFT	1,023	1,024	0,063
		BJ	1,000	0,999	0,056
		PAR	1,001	1,000	0,056
	β_2	AFT	1,028	1,028	0,059
		BJ	1,000	1,000	0,053
		PAR	1,000	1,000	0,053

Tabella 5.4: Risultati per distribuzione NORMALE con $\sigma = 1$

			Distribuzione campionaria		
n		metodo	media	mediana	dev. std
100	β_1	AFT	1,023	1,036	0,294
		BJ	1,011	1,017	0,255
		PAR	1,010	1,015	0,255
	β_2	AFT	1,010	1,018	0,238
		BJ	0,996	0,993	0,213
		PAR	0,996	0,994	0,213
200	β_1	AFT	1,017	1,004	0,186
		BJ	0,998	0,993	0,166
		PAR	0,998	0,993	0,165
	β_2	AFT	1,024	1,017	0,200
		BJ	1,004	1,000	0,175
		PAR	1,004	1,001	0,175
500	β_1	AFT	0,992	0,989	0,130
		BJ	0,991	0,986	0,129
		PAR	0,994	0,992	0,105
	β_2	AFT	1,001	1,003	0,110
		BJ	1,002	1,004	0,122
		PAR	0,999	1,005	0,102

Tabella 5.5: Risultati per distribuzione LOGISTICA con $\sigma = 0.5$

			Distribuzione campionaria		
n		metodo	media	mediana	dev. std
100	β_1	AFT	1,030	1,037	0,255
		BJ	1,007	1,013	0,211
		PAR	1,006	1,007	0,205
	β_2	AFT	1,032	1,033	0,274
		BJ	1,003	1,005	0,236
		PAR	1,001	0,997	0,231
200	β_1	AFT	1,035	1,031	0,187
		BJ	1,003	0,998	0,155
		PAR	1,001	1,003	0,151
	β_2	AFT	1,029	1,025	0,166
		BJ	0,992	0,988	0,138
		PAR	0,991	0,993	0,134
500	β_1	AFT	1,040	1,043	0,114
		BJ	1,006	1,007	0,101
		PAR	1,006	1,005	0,096
	β_2	AFT	1,039	1,041	0,110
		BJ	0,997	0,995	0,095
		PAR	0,996	0,995	0,090

Tabella 5.6: Risultati per distribuzione LOGISTICA con $\sigma = 1$

			Distribuzione campionaria		
n		metodo	media	mediana	dev. std
100	β_1	AFT	1,032	1,013	0,503
		BJ	0,994	1,005	0,427
		PAR	0,992	1,001	0,409
	β_2	AFT	1,012	0,972	0,537
		BJ	0,980	0,980	0,457
		PAR	0,972	0,956	0,448
200	β_1	AFT	1,061	1,038	0,472
		BJ	1,003	1,007	0,321
		PAR	1,002	1,019	0,309
	β_2	AFT	1,084	1,048	0,386
		BJ	1,022	1,019	0,289
		PAR	1,017	1,024	0,284
500	β_1	AFT	1,029	1,043	0,191
		BJ	0,994	1,003	0,189
		PAR	0,992	0,997	0,186
	β_2	AFT	1,067	1,045	0,193
		BJ	0,991	0,988	0,196
		PAR	0,989	0,995	0,189

Le Tabelle 5.1 e 5.2 mostrano i risultati quando i tempi di sopravvivenza sono simulati da una distribuzione Gumbel con parametro di scala pari rispettivamente a 0.5 e a 1.

I risultati riportati nelle tabelle evidenziano che gli stimatori sono non di-

storti, infatti i valori delle medie della distribuzione campionaria sono tutti uguali ai veri valori dei parametri: $\beta_1 = \beta_2 = 1$.
Le mediane sono approssimativamente pari alle medie e questo suggerisce che non vi siano valori anomali.
Le deviazioni standard della distribuzione campionaria nel modello AFT (che ricordiamo è quello stimato attraverso il modello di regressione esteso implementato in questa tesi) sono sempre inferiori rispetto a quelle stimate dal metodo Buckley e James (BJ) e risultano, comunque, confrontabili rispetto al metodo parametrico di stima (PAR) che è considerato il *gold standard*. Si osserva, quindi, l'efficienza degli stimatori ottenuti con il metodo AFT rispetto a quelli ottenuti con il metodo BJ .
Infine, si osserva che all'aumentare di n le deviazioni standard della distribuzione campionaria nel modello AFT assumono valori piú bassi confermando ulteriormente la buona performance rispetto a BJ.

Le Tabelle 5.3 e 5.4 mostrano i risultati della simulazione da una Normale con parametro di scala pari rispettivamente a 0.5 e a 1, mentre le Tabelle 5.5 e 5.6 sono relative alle simulazioni da una Logistica con i due diversi parametri di scala.
I risultati riportati nelle tabelle evidenziano che gli stimatori sono ancora non distorti, ma a differenza delle Tabelle 5.1 e 5.2, qui il comportamento degli stimatori AFT non è migliore rispetto a BJ, ma sicuramente confrontabile; cosí come risulta confrontabile anche con le stime ottenute attraverso il metodo parametrico.

Capitolo 6

Applicazioni a dati reali

In questo capitolo vengono analizzati due differenti database, molto noti in letteratura, per verificare i risultati dell'approccio proposto in questa tesi per la stima dei modelli PH, AFT ed AH attraverso il modello di regressione esteso. In particolare si osserva che nel primo esempio è verificata l'assunzione di proporzionalitá dei rischi del modello PH, nel secondo esempio, invece tale assunzione non è verificata, pertanto risulta piú opportuno l'utilizzo di un metodo alternativo per l'analisi della sopravvivenza.

6.1 Applicazione 1: breast cancer

In questo paragrafo vengono presentati i risultati relativamente ai dati di uno studio condotto dal *Middlesex Hospital* sul cancro al seno. Lo scopo di questo studio è stato quello di valutare se un particolare marker istochimico poteva essere usato per predire il tempo di sopravvivenza delle donne che avevano un cancro al seno. A tale scopo, sono stati analizzati i dati relativi a 45 donne che avevano avuto una mastectomia per trattare il tumore tra

Gennaio 1969 e Dicembre 1971. Le sezioni dei tumori venivano trattate con il marker istochimico (HPA: Helix Pomatia Agglutinin) e ciascun tumore veniva successivamente classificato come positivamente o negativamente colorato. La colorazione positiva corrisponde ad un tumore con metastasi. Le variabili utilizzate per l'analisi sono:

- *tempo*: esprime il tempo di sopravvivenza (in mesi) dall'intervento chirurgico;

- *status*: assume valore 0 per i censurati e valore 1 per i deceduti;

- *gruppo*: assume valore 0 per i tumori con colorazione negativa e valore 1 per i tumori con colorazione positiva.

L'obiettivo è valutare se c'è una differenza significativa nella sopravvivenza per i due gruppi di donne.
Iniziamo l'analisi dei dati considerando, dapprima, un modello di Cox a rischi proporzionali. In particolare inseriamo nel modello l'esplicativa gruppo.

Per verificare l'assunzione di proporzionalitá dei rischi in un modello PH utilizziamo un metodo grafico che utilizza il logaritmo della funzione hazard cumulata stimata per ciascun gruppo rispetto al tempo di sopravvivenza. Funzioni parallele indicano che l'assunzione è verificata. La Figura 6.1 mostra che puó essere considerata valida l'assunzione di proporzionalitá dei rischi.

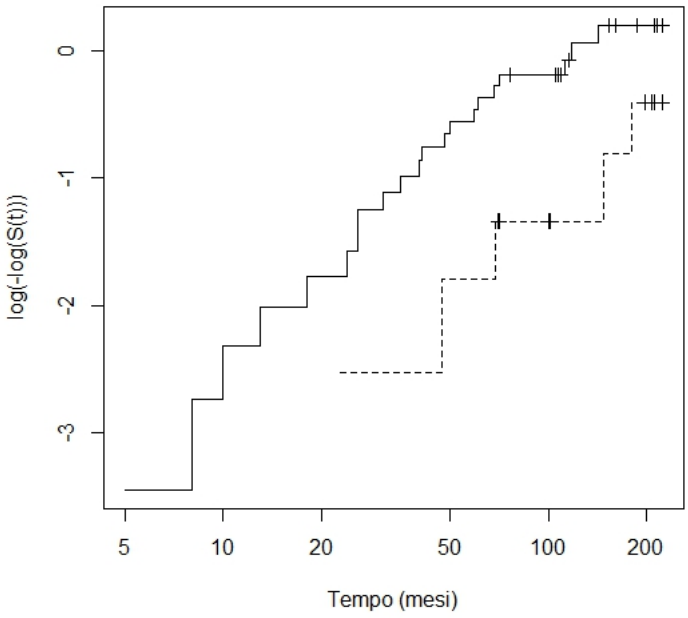

Figura 6.1: Valutazione grafica per la proporzionalitá dei rischi: dati breast cancer

Vediamo i risultati dell'approccio proposto in questa tesi per la stima dei modelli PH, AFT, AH ed EH attraverso il modello di regressione esteso. La Tabella 6.1 riporta le devianze basate sul modello di Poisson per i diversi modelli di sopravvivenza e i gradi di libertá per ogni modello stimato. Si noti che quest'ultimi sono dati da 1 per il coefficiente del trattamento e 5 per il termine $\lambda_0(t)$. Si osservi, inoltre che nel modello EH si ha un grado di libertá in piú corrispondente ad un ulteriore coefficiente presente nel modello (4.1). La Tabella 6.1 suggerisce che è preferibile un modello

PH.

Tabella 6.1: Confronto tra modelli - Devianze basate sul modello di Poisson: dati breast cancer

Modello	dev	df
PH	214,5	6
AFT	214,6	6
AH	219,1	6
EH	212,8	7

Dal confronto tra i modelli e dall'analisi grafica della proporzionalitá dei rischi, si sceglie di adattare un modello PH ai dati in studio. Scelto il modello, si procede con la stima dei parametri ed, in particolare, si riporta la Tabella 6.2 che presenta i risultati ottenuti utilizzando diversi strumenti per modellare $\log(\lambda_0(t))$, che sono i polinomi, le *Natural Spline* e le *B-Spline*.

Tabella 6.2: Stima del coefficiente della variabile gruppo per il modello PH: dati breast cancer

	Stime	SE
Polinomi	0,909	0,50
B-Spline	0,908	0,50
Natural Spline	-0,907	0,50

I risultati riportati nella Tabella 6.2 mostrano che le differenze sulle stime e sugli errori standard sono trascurabili, quindi si puó scegliere indifferentemente tra i tre strumenti proposti per modellare la funzione *ha-*

zard baseline. Gli errori standard sono ottenuti come inverso della matrice Hessiana ottenuta numericamente e valutata nella soluzione ottenuta.

6.2 Applicazione 2: gastric cancer

In questo paragrafo vengono presentati i risultati relativamente ai dati di uno studio condotto dal *Gastrointestinal Tumor Study Group* (1982), allo scopo di confrontare due differenti trattamenti del cancro allo stomaco non operabile che consistono di sola chemioterapia e di una combinazione tra chemioterapia e radioterapia. In particolare, lo studio viene condotto su 90 pazienti. Di essi, 45 vengono assegnati al trattamento con sola chemioterapia e 45 al trattamento combinato.
Le variabili utilizzate per l'analisi sono:

- *tempo*: esprime il tempo di sopravvivenza (in giorni) dal trattamento;
- *status*: assume valore 0 per i censurati e valore 1 per i deceduti;
- *trattamento*: assume valore 0 per il trattamento con chemioterapia e valore 1 per il trattamento combinato.

L'obiettivo è valutare se c'è una differenza significativa nella sopravvivenza dei pazienti per i due gruppi di trattamento.
Iniziamo l'analisi dei dati considerando, dapprima, un modello di Cox a rischi proporzionali. In particolare inseriamo nel modello l'esplicativa trattamento.

Per verificare l'assunzione di proporzionalitá dei rischi in un modello PH utilizziamo, come per l'esempio precedente, il metodo grafico che utilizza il logaritmo della funzione hazard cumulata stimata per ciascun gruppo rispetto al tempo di sopravvivenza. La Figura 6.2 mostra che le funzioni non sono parallele, pertanto non puó essere considerata valida l'assunzione di proporzionalitá dei rischi.

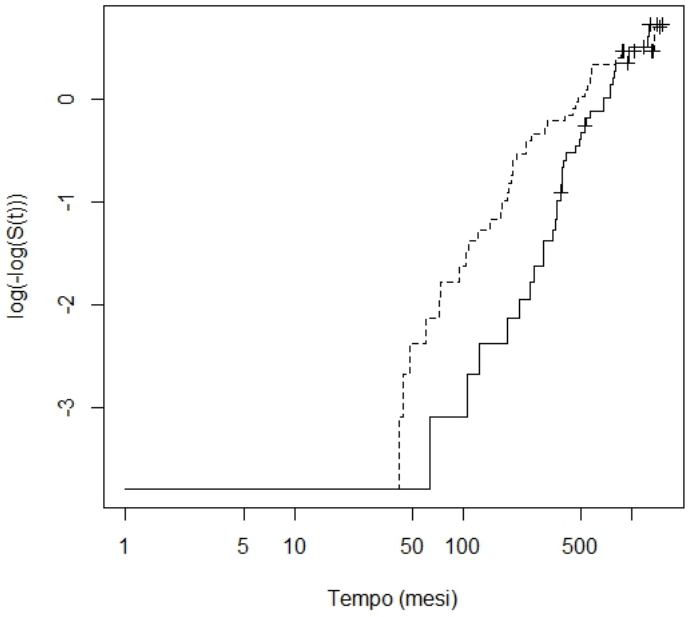

Figura 6.2: Valutazione grafica per la proporzionalitá dei rischi: dati gastric cancer

Consideriamo l'approccio proposto in questa tesi per la stima dei modelli PH, AFT, AH ed EH attraverso il modello di regressione esteso. Le

stime dei parametri vengono riportate nella Tabella 6.3. Come nell'applicazione precedente, la Tabella 6.3 riporta le devianze basate sul modello di Poisson per i diversi modelli di sopravvivenza e i gradi di libertá per ogni modello stimato. Anche in quest'analisi i gradi di libertá sono dati da 1 per il coefficiente del trattamento e 5 per il termine $\lambda_0(t)$. Nel modello EH si ha un grado di libertá in piú (4.1).

Tabella 6.3: Confronto tra modelli - Devianze basate sul modello di Poisson: dati gastric cancer

Modello	dev	df
PH	542,1	6
AFT	537,2	6
AH	543,2	6
EH	533,5	7

L'analisi grafica della proporzionalitá dei rischi mostra che non è consigliabile adattare ai dati in studio un modello PH perché l'assunzione di proporzionalitá dei rischi sembra non essere soddisfatta; inoltre, dal confronto tra le devianze e i gradi di libertá di ogni modello stimato, si evince che è preferibile scegliere il modello AFT. La Tabella 6.3 mostra, infatti, che il modello EH ha una devianza non molto piú piccola di quella del modello AFT e tale differenza risulta statisticamente non significativa al 5%. Pertanto si sceglie il modello piú semplice (in termini di gradi di libertá), ovvero il modello AFT.
Scelto il modello, si procede con la stima dei parametri ed, in particolare, si riporta la Tabella 6.4 che presenta i risultati ottenuti utilizzando diversi strumenti per modellare $\log(\lambda_0(t))$, che sono i polinomi, le *Natural Spline* e

le *B-Spline*.

Tabella 6.4: Stima del coefficiente della variabile trattamento per il modello AFT: dati gastric cancer

	Stime	SE
Polinomi	0,406	0,236
B-Spline	0,405	0,237
Natural Spline	0,403	0,237

I risultati riportati nella Tabella 6.4 mostrano che le differenze sulle stime e sugli errori standard sono trascurabili quindi la funzione *hazard baseline* puó essere stimata indifferentemente con uno dei tre strumenti proposti. Come nell'applicazione precedente, gli errori standard sono ottenuti come inverso della matrice Hessiana ottenuta numericamente e valutata nella soluzione ottenuta.

Conclusioni

In questa tesi è stato proposto un nuovo impianto modellistico per l'analisi di dati di sopravvivenza attraverso un approccio semiparametrico, ovvero senza specificare la distribuzione dei tempi, ma assumendo una determinata forma funzionale con le esplicative. È stato discusso un modello di regressione hazard esteso che comprende come casi particolari il modello di Cox a rischi proporzionali, il modello *Accelerated Failure Time* e, infine, il modello *Accelerated Hazard*. Il *framework* proposto si basa su una regressione di Poisson e la stima si ottiene attraverso un processo iterativo. La trattazione si è concentrata sull'algoritmo di stima e alcune simulazioni hanno mostrato che l'algoritmo porta a stimatori con comportamento soddisfacente ed in linea con altri competitori.

Ci sono almeno due aspetti che meriterebbero di essere approfonditi: il calcolo delle varianze degli stimatori ed una valutazione piú approfondita sull'utilizzo del modello EH a fini discriminativi tra i diversi modelli piú semplici (ad es., PH o AFT).

Bibliografia

Aitkin, M. and Clayton, D. (1980). The fitting of exponential, weibull and extreme value distributions to complex censored survival data using glim. *Applied Statistics*, **29**, 156–163.

Allison, P. (2010). *Survival Analysis Using SAS System: A Practical Guide*. Cary NC: SAS Institute.

Buckley, J. and James, I. (1979). Linear regression with censored data. *Biometrika*, **66**(3), 429–436.

Carstensen, B. (2005). Demography and epidemiology: Practical use of the lexis diagram in the computer age. In *Annual Meeting of Finnish Statistical Society*, pages 1–30.

Chen, Y. (2001). Accelerated hazards regression model and its adequacy for censored survival data. *Biometrics*, **57**, 853–860.

Chen, Y. and Jewell, N. (2001). On a general class of semiparametric hazards regression models. *Biometrika*, **88**(3), 687–702.

Chen, Y. and Wang, M. (2000a). Analysis of accelerated hazards models. *Journal of the American Statistical Association*, **95**, 608–618.

Chen, Y. and Wang, M. (2000b). Estimating a treatment effect with the accelerated hazards models. *Controlled Clinical Trials*, **21**, 369–380.

Chen, Y., Hanson, T., and Zhang, J. (2014). Accelerated hazards model based on parametric families generalized with bernstein polynomials. *Biometrics*, **70**, 192–201.

Collett, D. (2003). *Modelling Survival Data in Medical Research*. Chapman & Hall, London.

Cox, D. (1972). Regression models and life tables. *Journal of Royal Statistical Society: Series B*, **34**(2), 187–220.

Cox, D. and Oakes, D. (1984). *Analysis of Survival Data*. Chapman & Hall, London.

Etezadi-Amoli, J. and Ciampi, A. (1987). Extended hazard regression for censored survival data with covariates: A spline approximation for the baseline hazard function. *Biometrics*, **43**, 181–192.

Fahrmeir, L. and Tutz, G. (1994). *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer, New York.

Fleming, T. R. and Harrington, D. (2005). *Counting Processes and Survival Analysis*. Wiley.

Fox, J. (2002). Cox proportional-hazards regression for survival data. Appendix to An R and S-PLUS Companion to Applied Regression.

Glasser, M. (1967). Exponential survival with covariance. *Journal of the American Statistical Association*, **62**, 561–568.

Hosmer, D. and Lemeshow, S. (1999). *Applied Survival Analysis: Regression Modeling of Time to Event Data*. Wiley, New York.

Jin, Z., Lin, D. Y., and Ying, Z. (2003). Rank-based inference for the accelerated failure time model. *Biometrika*, **90**, 341–353.

Jin, Z., Lin, D. Y., and Ying, Z. (2006). On least-squares regression with censored data. *Biometrika*, **93**, 147–161.

Kalbfleisch, J. and Prentice, R. (2002). *The Statistical Analysis of Failure Time Data*. Wiley.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, **53**, 457–481.

Klein, J. P. and Moeschberger, M. (2003). *Survival Analysis Techniques for Censored and Truncated Data*. Springer.

Kleinbaum, D. and Klein, M. (2005). *Survival Analysis: A self-learning text*. Springer.

Lawless, J. (2003). *Statistical Models and Methods for Lifetime Data*. Wiley.

Lee, E. T. and Wang, J. (2003). *Statistical Method for Survival Data Analysis*. Wiley.

Martinussen, T. and Scheike, T. (2006). *Dynamic Regression Models for Survival Data*. Springer.

Marubini, E. and Valsecchi, M. (1995). *Analysing Survival Data from Clinical Trials and Observational Studies*. Wiley.

McCullagh, P. and Nelder, J. (1988). *Generalized Linear Models*. Chapman & Hall, 2nd edition.

Miller, R. and Halpern, J. (1982). Regression with censored data. *Biometrika*, **69**, 521–531.

Orbe, J. and Nunez-Anton, V. (2006). Alternative approaches to study lifetime data under different scenarios: from the ph to the modified semiparametric aft model. *Computational Statistics and Data Analysis*, **50**, 1565–1582.

Orbe, J., Ferreira, E., and Nunez-Anton, V. (2002). Comparing proportional hazards and accelerated failure time models for survival analysis. *Statistics in Medicine*, **21**, 3493–3510.

Perperoglou, A. (2011). Fitting survival data with penalized poisson regression. *Stat Methods Appl*, **20**, 451–462.

Rossi, P., Berk, R., and Lenihan, K. (1980). *Money, Work and Crime: Some Experimental Results*. New York: Academic Press.

Selvin, S. (2008). *Survival Analysis for Epidemiologic and Medical Research*. Cambridge University Press.

Singer, J. and Willet, J. (2003). *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*. Oxford University Press.

Tableman, M. and Kim, J. (2004). *Survival Analysis Using S*. Chapman & Hall.

Tong, X., Zhu, L., Leng, C., Leisenring, W., and Robison, L. (2013). A general semiparametric hazards regression model: efficient estimation and structure selection. *Statistics in Medicine*, **32**, 4980–4994.

Tseng, Y. and Shu, K. (2011). Efficient estimation for a semiparametric extended hazards model. *Communications in Statistics - Simulation and Computation*, **40**, 258–273.

Tseng, Y., Hsu, K., and Yang, Y. (2014). A semiparametric extended hazard regression model with time-dependent covariates. *Journal of Nonparametric Statistics*, **26**(1), 115–128.

Whitehead, J. (1980). Fitting cox regression model to survival data using glim. *Applied Statistics*, **29**(3), 268–275.

Zeng, D. and Lin, D. (2007). Efficient estimation for the accelerated failure time model. *Journal of the American Statistical Association*, **102**(480), 1387–1396.

Zhang, J., Peng, Y., and Zhao, O. (2011). A new semiparametric estimation method for accelerated hazard model. *Biometrics*, **67**, 1352–1360.